Geneious 2.0

Biomatters Ltd

August 2, 2006

Contents

1 Getting Started **5**

1.1 Downloading & Installing Geneious 5

1.2 Using Geneious for the first time 7

1.3 Troubleshooting 10

2 Data handling **13**

2.1 Importing and exporting data 13

2.2 Acquiring data from public databases 18

2.3 Searching databases and local documents using Geneious 22

2.4 Agents 26

2.5 Storing data in Geneious 29

2.6 Panels in Geneious 32

2.7 Document Viewers in Geneious 37

2.8 Preferences 48

3 Data Processing **51**

3.1 Data Filtering 51

3.2 Literature 56

3.3 Sequence data 57

3.4 Dotplots 57

3.5 Pairwise sequence alignments 58

3.6 Multiple sequence alignments 60
3.7 Sequence alignment using ClustalW (*pro* only) 62
3.8 Building Phylogenetic trees 63
3.9 BLAST 67
3.10 Results of analysis 67

4 Collaboration (*pro* only) **69**
4.1 Managing Your Accounts 69
4.2 Managing Your Contacts 72
4.3 Sharing Documents 74
4.4 Searching Contacts 74

Chapter 1

Getting Started

By the end if this chapter, you should:

- Be able to download, install, and upgrade Geneious
- Be able to import Notes and Documents from earlier versions of Geneious
- Be familiar with the layout of Geneious
- Be able to solve connection problems.

1.1 Downloading & Installing Geneious

Geneious is free software that can be downloaded from http://www.geneious.com.

The panel on the right hand side of the Geneious home page (Figure 1.1) allows users to download Geneious onto three operating systems - Windows, Mac OS X or Linux. Make sure your system meets the requirements before downloading Geneious.

Click on the operating system to take you to the "Download Geneious" page. Follow the instructions on the page to download and save Geneious. It is often easiest to save the program to your desktop.

Once Geneious is saved, double left-click on the Geneious icon to start installing the program. While this is happening, you will be prompted for a location to install Geneious. Please check that you are satisfied with the location before continuing.

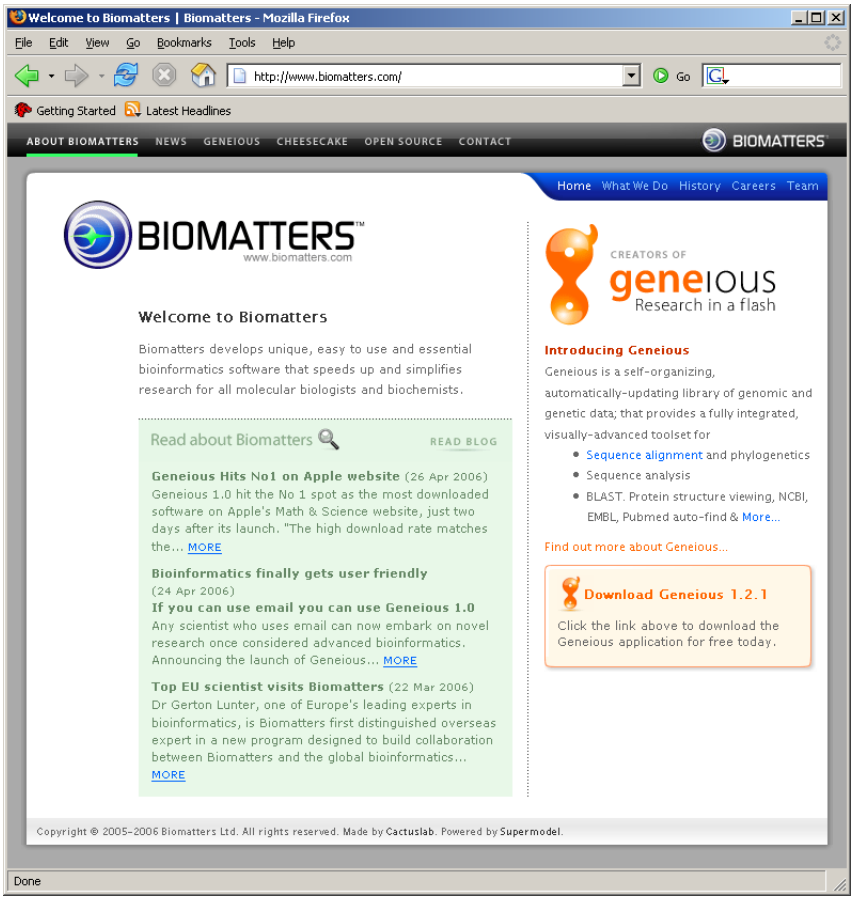

Welcome to Biomatters | Biomatters - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.biomatters.com/
Go
Getting Started
Latest Headlines
ABOUT BIOMATTERS NEWS GENEIOUS CHEESECAKE OPEN SOURCE CONTACT
BIOMATTERS
Home What We Do History Careers Team
BIOMATTERS
www.biomatters.com
Welcome to Biomatters
Biomatters develops unique, easy to use and essential bioinformatics software that speeds up and simplifies research for all molecular biologists and biochemists.
Read about Biomatters
READ BLOG
Geneious Hits No1 on Apple website (26 Apr 2006)
Geneious 1.0 hit the No 1 spot as the most downloaded software on Apple's Math & Science website, just two days after its launch. "The high download rate matches the... MORE
Bioinformatics finally gets user friendly
(24 Apr 2006)
If you can use email you can use Geneious 1.0
Any scientist who uses email can now embark on novel research once considered advanced bioinformatics. Announcing the launch of Geneious... MORE
Top EU scientist visits Biomatters (22 Mar 2006)
Dr Gerton Lunter, one of Europe's leading experts in bioinformatics, is Biomatters first distinguished overseas expert in a new program designed to build collaboration between Biomatters and the global bioinformatics... MORE
CREATORS OF
geneious
Research in a flash
Introducing Geneious
Geneious is a self-organizing, automatically-updating library of genomic and genetic data; that provides a fully integrated, visually-advanced toolset for
Sequence alignment and phylogenetics
Sequence analysis
BLAST. Protein structure viewing, NCBI, EMBL, Pubmed auto-find & More...
Find out more about Geneious...
Download Geneious 1.2.1
Click the link above to download the Geneious application for free today.
Copyright © 2005-2006 Biomatters Ltd. All rights reserved. Made by Cactuslab. Powered by Supermodel.
Done

Figure 1.1: The Geneious home page

1.1.1 Upgrading to new versions

All versions of Geneious from 0.9d onwards will automatically self-update and retain all of your data. If you are upgrading from 0.9b or earlier, please see the section: "Upgrading from Version 0.9b or earlier".

1.2 Using Geneious for the first time

Figure 1.2 displays the Geneious window when opened for the first time. There are six main panels.

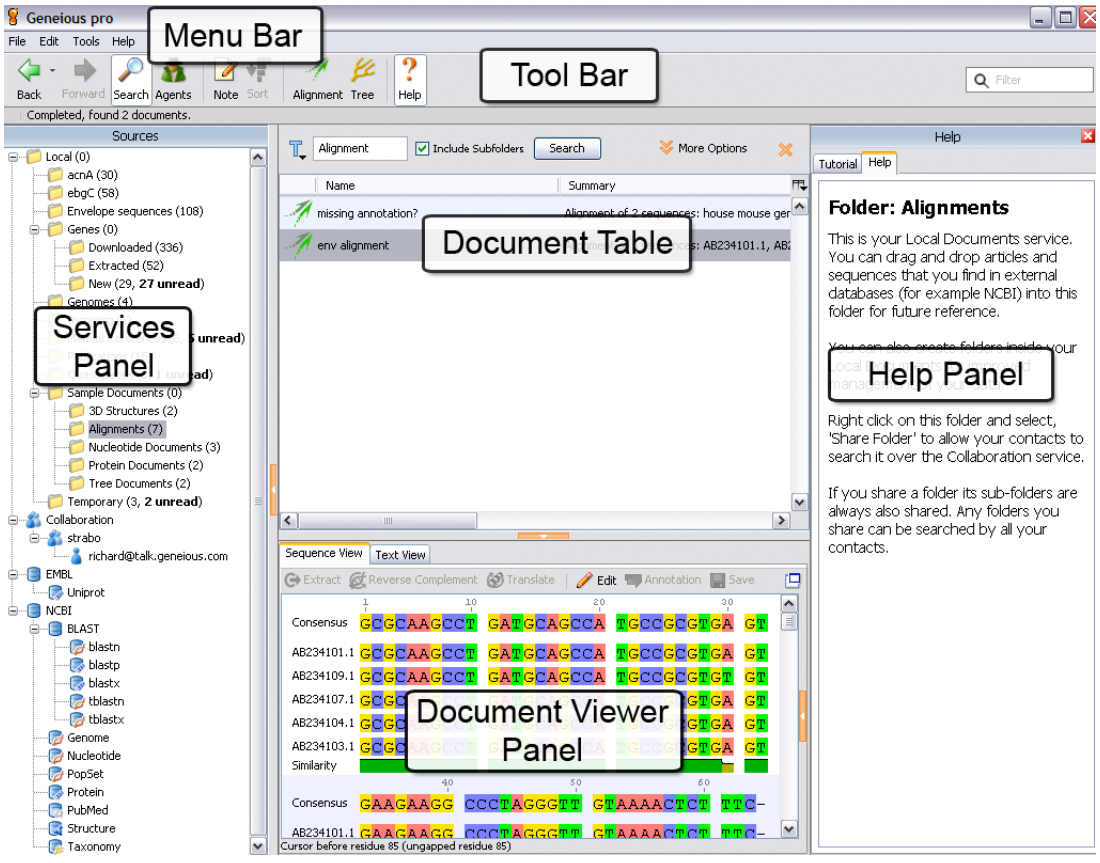

Menu Bar
Tool Bar
Services Panel
Document Table
Document Viewer Panel
Help Panel
Folder: Alignments
This is your Local Documents service. You can drag and drop articles and sequences that you find in external databases (for example NCBI) into this folder for future reference.
Right click on this folder and select, 'Share Folder' to allow your contacts to search it over the Collaboration service.
If you share a folder its sub-folders are always also shared. Any folders you share can be searched by all your contacts.

Figure 1.2: The main window in Geneious

1.2.1 The Services Panel

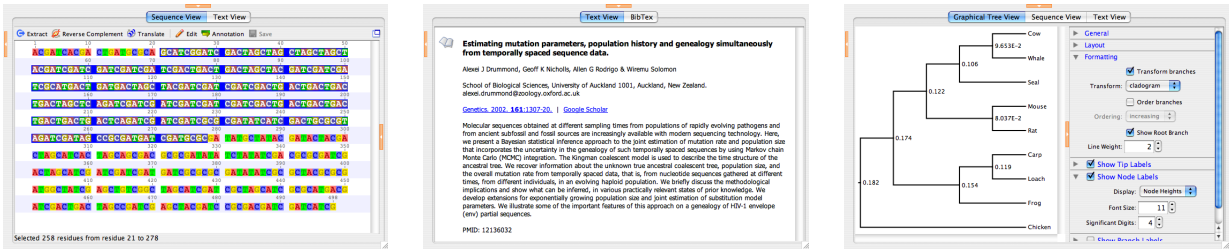
The Services Panel contains the services and functions Geneious offers. These include your local documents (including sample documents), EMBL and NCBI links and Collaboration. All these services will be described in detail later in the manual.

1.2.2 The Document Table

The Document Table displays summaries of downloaded data such as DNA sequences, protein sequences, journal articles, sequence alignments, and trees. By clicking on the search icon you can search data for text or by sequence similarity (BLAST). Data can also be filtered using the "Filter" box located at the right side of the toolbar.

1.2.3 The Document Viewer Panel

The Document Viewer Panel is where sequences, alignments, trees, and journal article abstracts can be shown graphically or as plain text. This panel also offers various options while visualizing protein and nucleotide sequences. These options include zooming, color and layout selection, and annotations. When viewing trees, there are additional options for branch and leaf labeling, and controlling tree layout. When viewing journal articles, this panel includes a direct link to Google Scholar. All these options are displayed on the right-hand side of the panel (Figure 1.3).



(a) Nucleotide sequence (b) Journal Article (c) Phylogenetic tree

Figure 1.3: Three document viewers

1.2.4 The Help Panel

The Help Panel includes a tutorial. If you are new to Geneious, we recommend working through the tutorial first. Most help is automatic. A short help message pops up when any of the tasks on the Services panel are clicked. This panel can be closed at any time by clicking

the "X" symbol in the top right-hand corner. If you want to display the help panel, click on "Help" in the Menu Bar and select the small box beside "Help". The Help Panel will reappear (Figure 1.4).

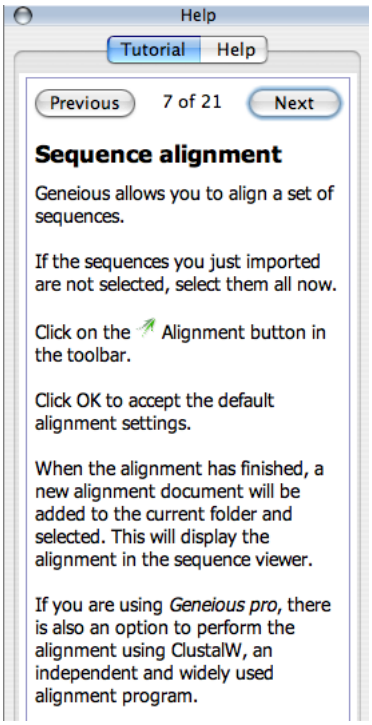

Help
Tutorial
Help
Previous
7 of 21
Next
Sequence alignment
Geneious allows you to align a set of sequences.
If the sequences you just imported are not selected, select them all now.
Click on the Alignment button in the toolbar.
Click OK to accept the default alignment settings.
When the alignment has finished, a new alignment document will be added to the current folder and selected. This will display the alignment in the sequence viewer.
If you are using Geneious pro, there is also an option to perform the alignment using ClustalW, an independent and widely used alignment program.

Figure 1.4: The Help Panel

1.2.5 The Toolbar

The Toolbar displays nine large icons : "Back", "Forward", "Search", "Agents", "Add Note", "Sort", "Alignment", "Tree", "Help" – and a drop down arrow in between the "Back" and "Forward" icons providing quick access to previous views. The Toolbar also has a "Filter" box at the far right. The "Search" icon is available when local documents are selected. The "Add Note" icon is available when one or more documents are selected. The "Sort" icon is available when exactly one sequence is selected. Sorting is done by "BLAST-ing" (2.5.2) the sequence against all other sequences in the folder and ordering according based on E-values.

The Alignment icon only becomes available when two or more protein, nucleotide sequences or alignments are selected in the Document Table. The Tree icon becomes available when an alignment or a set of sequences is selected. The toolbar also allows the user to hide icons when they are not in use (Figure 1.5). To hide the icons, right-click anywhere on the Toolbar to see a list of the icons. Then, scroll down to the icon you want to hide and left-click. The tick (check)

beside the icon will disappear and the icon will disappear from the Toolbar. To make the icons reappear, follow the above steps in reverse.

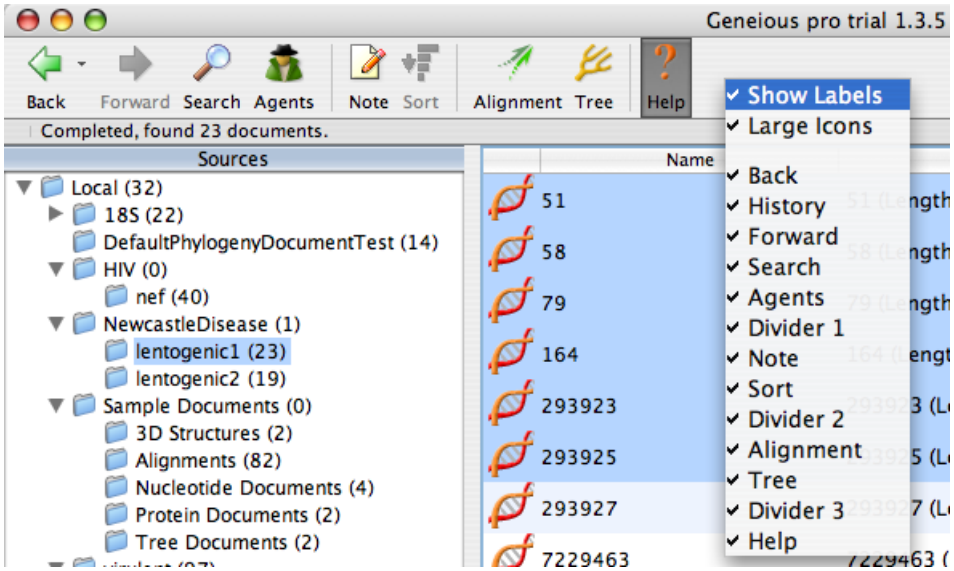


Figure 1.5: The Toolbar

1.2.6 The Menu Bar

The Menu Bar has five main menus “File”, “Edit”, “Tools”, “Collaboration” and “Help”. These are discussed in more detail at the end of Chapter 2.

1.3 Troubleshooting

1.3.1 Geneious won’t start

Geneious has some minimum system requirements. It is compatible with the three most common operating systems: Windows, Mac, and Linux. Check that you have one of the following OS versions before you launch Geneious:

Operating System	System requirements
Windows	2000/XP
Mac OS X	10.4
Unix/Linux	

Geneious also needs Java 1.5 to run. If you do not have this on your system already, please download it from http://www.geneious.com.

On the download page, select the "Includes Java 1.5" option. This involves downloading a larger file.

If you are a Mac user, and have OS X 10.4 or later, you will have to download Java 1.5 from http://www.apple.com/support/downloads/java2se50release3.html.

1.3.2 I get a connection error when trying to search using NCBI or EMBL

If the message reads, "Check your connection settings", there is a problem with your Internet connection. Make sure you are still connected to the Internet. Both Dial-up and Broadband can disconnect. If you are connected, then the error message indicates you are behind a proxy server and Geneious has been unable to detect you proxy settings automatically. You can fix this problem:

1. Check the browser you are using. These instructions are for Explorer, Safari, and Firefox.
2. Open up your default browser.
3. Use the steps in Figure 1.6 for each browser to find the connection settings.
4. Now go into Geneious and select "Preferences". There are two ways to do this.
   - *Shortcut keys*. Ctrl+Shift+P (Windows/Linux), Command+Shift+P (Mac OS X).
   - *Tools Menu → Preferences*.
5. This opens the Preferences. Click on the "General" tab. There are four options in the drop-down options under "Connection settings" (Figure 1.7):
   - *Use direct connection*. Use this setting when no proxy settings are required.
   - *Use browser connection settings*. This allows Geneious to automatically import the proxy settings.
   - *Use proxy server*. This enables two text fields : Proxy host and Proxy port. This information is in your browser's connection settings. Please see step 3.
   - *Use auto config file*. This enables one text field called "Config file location". These details can also be found in your browser's settings.
6. Finally, set the proxy settings under the General tab to match those in your browser.

Note. If you are using any other browser, and cannot find the proxy settings, please email us at support@geneious.com.

Figure 1.6: Checking browser settings

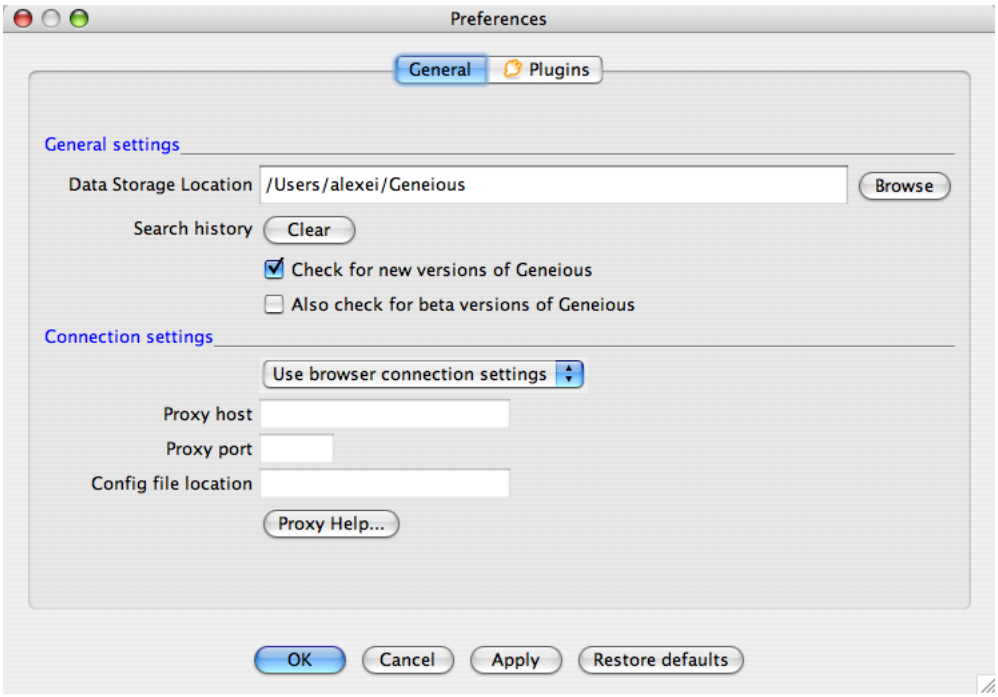


Figure 1.7: General Preferences

Chapter 2

Data handling

Geneious is a one-stop-shop for handling and managing your bioinformatic data. This chapter summarizes the different ways you can use Geneious to acquire, update, organize, store, and view your data.

By the end of this chapter, you should be able to:

- Import/Export data from various sources
- Organize your data into easily accessible folders
- Automatically update your data
- Know the purpose of each panel in Geneious
- Customize Geneious to meet your needs.

2.1 Importing and exporting data

Geneious is able to import raw data from different applications and export the results in a range of formats. If you are new to bioinformatics, please take the time to familiarize yourself with this chapter as there are a number of formats to be aware of.

2.1.1 Importing data from the hard drive to your Local folders

To import files from your hard disk, click "File" → "Import" → "From file". This will open up a file dialog. Select one or more files and click "Import". If Geneious automatic file format detection fails, select the file type you wish to import (Figure 2.1) before selecting the name of the file(s). The different file types are described in detail in the next section..

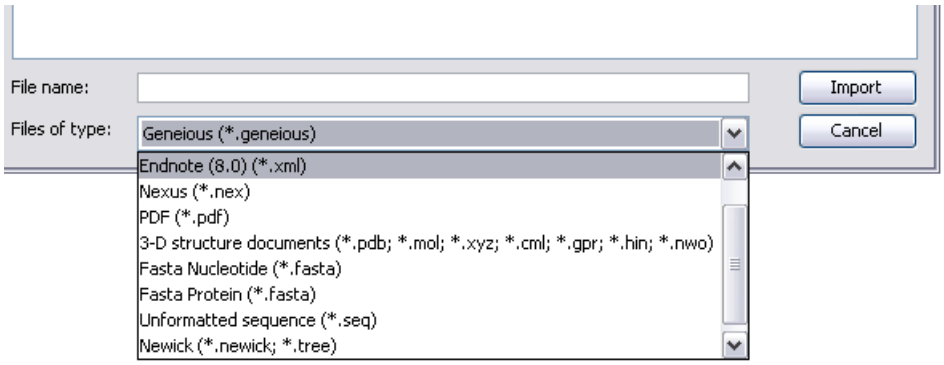


Figure 2.1: File import options

2.1.2 Data input formats

Geneious version 2.0 can handle all the following file formats.

Format	Extensions	Data types	Common sources
Clustal	*.aln	Alignments	ClustalX
Endnote (8.0) XML	*.xml	Journal article references	Endnote, Journal article websites
Fasta	*.fasta, *.fas, etc.	Sequences, alignments	PAUP*, ClustalX, BLAST, FASTA
Geneious	*.xml, *.geneious	Preferences, databases	Geneious
Newick	*.tre, *.tree, etc.	Phylogenetic trees	PHYLIP, Tree-Puzzle, PAUP*, ClustalX
Nexus	*.nxs, *.nex	Trees, Alignments	PAUP*, Mesquite, MrBayes and MacClade
PileUp	*.msf	Alignments	pileup (gcg)
Rich Sequence Format	*.rsf	Sequences, alignments	GCGs NetFetch
DNA Strider	*.str	sequence	DNA Strider (Mac program), ApE
PDB	*.pdb	3D Protein structures	SP3, SP2, SPARKS, Protein Data Bank
PDF	*.pdf	Documents, presentations	Adobe Writer, LaTeX, Miktex
Raw sequence text	*.seq	Sequences	Any file that contains only a sequence
Sequence Chromatograms	*.ab1, *.scf	Raw Sequencing trace and sequence	Sequencing machines

CLUSTAL format

The Clustal format is used by ClustalW [23] and ClustalX [22], two well known multiple sequence alignment programs.

Clustal format files are used to store multiple sequence alignments and contain the word clustal at the beginning. An example Clustal file:

```
CLUSTAL W (1.74) multiple sequence alignment
```

```
seq1 ----------------------KSKERYKDENGGNYFQLREDWWDANRETVWKAITCNA
seq2 ---------------YEGLTTANGXKEYYQDKNGGNFFKLREDWWTANRETVWKAITCGA
seq3 ----KRIYKKIFKEIHSGLSTKNGVKDRYQN-DGDNYFQLREDWWTANRSTVWKALTCSD
seq4 ----------------------SQRHYKD-DGGNYFQLREDWWTANRHTVWEAITCSA
seq5 --------------------NVAALKTRYEK-DGQNFYQLREDWWTANRATIWEAITCSA
seq6 ------FSKNIX--QIEELQDEWLLEARYKD--TDNYYELREHWWTENRHTVWEALTCEA
seq7 -------------------------------------------------KELWEALTCSR

seq1 --GGGKYFRNTCDG--GQNPTETQNNCRCIG----------ATVPTYFDYVPQYLRWSDE
seq2 P-GDASYFHATCDSGDGRGGAQAPHKCRCDG---------ANVVPTYFDYVPQFLRWPEE
seq3 KLSNASYFRATC--SDGQSGAQANNYCRCNGDKPDDDKP-NTDPPTYFDYVPQYLRWSEE
seq4 DKGNA-YFRRTCNSADGKSQSQARNQCRC---KDENGKN-ADQVPTYFDYVPQYLRWSEE
seq5 DKGNA-YFRATCNSADGKSQSQARNQCRC---KDENGXN-ADQVPTYFDYVPQYLRWSEE
seq6 P-GNAQYFRNACS----EGKTATKGKCRCISGDP----------PTYFDYVPQYLRWSEE
seq7 P-KGANYFVYKLD-----RPKFSSDRCGHNYNGDP---------LTNLDYVPQYLRWSDE
```

EndNote 8.0 XML format

EndNote is a popular reference and bibliography manager. EndNote lets you search for journal articles online, import citations, perform searches on your own notes, and insert references into documents. It also generates a bibliography in different styles. Geneious can interoperate with EndNote using Endnote's XML (Extensible Markup Language) file format to export and import its files.

FASTA format

The FASTA file format is commonly used by many programs and tools, including BLAST [1], T-Coffee [16] and ClustalX [22]. Each sequence in a FASTA file has a header line beginning with a ">" followed by a number of lines containing the raw protein or DNA sequence data. The sequence data may span multiple lines and these sequence may contain gap characters. An empty line may or may not separate consecutive sequences. Here is an example of three sequences in FASTA format (DNA, Protein, Aligned DNA):

```
>Orangutan
ATGGCTTGTGGTCTGGTCGCCAGCAACCTGAATCTCAAACCTGGAGAGTGCCTTCGAGTG

>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQK

>Chicken
```

```
CTACCCCCCTAAAACACTTTGAAGCCTGATCCTCACTA------------------CTGT
CATCTTAA
```

Geneious format

The Geneious format can be used to store all your local documents, note types and program preferences. A file in Geneious format will usually have a `.geneious` extension or a `.xml` extension. This format is useful for sharing documents with other Geneious users and backing up your Geneious data.

Newick format

The Newick format is commonly used to represent phylogenetic trees (such as those inferred from multiple sequence alignments). Newick trees use pairs of parentheses to group related taxa, separated by a comma (,). Some trees include numbers (branch lengths) that indicate the distance on the evolutionary tree from that taxa to its most recent ancestor. If these branch lengths are present they are prefixed with a colon (:). The Newick format is produced by programs such as PHYLIP, PAUP*, ClustalW [23], ClustalX [22], Tree-Puzzle [7] and PROTML. Geneious is also able to read trees in Newick format and display them in the visualization window. It also gives you a number of display options including tree types, branch lengths, and labels.

Nexus format

The Nexus format [12] was designed to standardize the exchange of phylogenetic data, including sequences, trees, distance matrices and so on. The format is composed of a number of blocks such as TAXA, TREES and CHARACTERS. Each block contains pre-defined fields. Geneious imports and exports files in Nexus format, and can process the information stored in them for analysis.

PileUp format

The PileUp format is used by the pileup program, a part of the Genetics Computer Group (GCG) Wisconsin Package.

Rich Sequence format

RSF (Rich Sequence Format) files contain one or more sequences that may or may not be related. In addition to the sequence data, each sequence can be annotated with descriptive sequence

information.

DNA Strider

Sequence files generated by the Mac program DNA Strider, containing one Nucleotide or Protein sequence.

PDB format

Protein Databank files contain a list of XYZ co-ordinates that describe the position of atoms in a protein. These are then used to generate a 3D model which is usually viewed with Rasmol or SPDB viewer. Geneious can read PDB format files and display an interactive 3D view of the protein structure, including support for displaying the protein's secondary structure when the appropriate information is available.

PDF format

PDF stands for Portable Document Format and is developed and distributed by Adobe Systems (http://www.adobe.com/). It contains the entire description of a document including text, fonts, graphics, colors, links and images. The advantage of PDF files is that they look the same regardless of the software used to create them. Some word processors are able to export a document into PDF format. Alternatively, Adobe Writer can be used. Currently, you can use Geneious to read, store and open PDF files and future versions will have more options for storing and manipulating PDF.

Sequence Chromatograms

Sequence chromatogram documents contain the results of a sequencing run (the trace) and a guess at the sequence data (base calling).

Informally, the trace is a graph showing the concentration of each nucleotide against sequence positions. Base calling software detects peaks in the four traces and assigns the most probable base at more or less even intervals.

2.1.3 Where does my imported data go?

The above formats can be all imported into Geneious from files on your hard drive. Geneious also enables you to download certain types of documents directly from public databases such

as NCBI and EMBL. The method used to retrieve a particular piece of data will determine where in Geneious it is stored.

Data imported from your hard drive. This is imported directly into the currently selected local folder within Geneious. If no folder is selected, Geneious open a dialog which lets you specify a folder.

Data from an NCBI/EMBL/Contacts search. Data downloaded from public databases within Geneious will appear in the Document Table and can be dragged from there into a local folder of your choice.

Important: if you don't drag the documents from a database search into your local folders the results will be lost when Geneious is closed.

2.1.4 Data output formats

Each data type has several export options. Any set of documents may be exported in Geneious native format:

Data type	Export format options
DNA sequence	FASTA, Geneious
Amino acid sequence	FASTA, Geneious
Protein 3D structure	PDB, FASTA, Geneious
Multiple sequence alignment	FASTA, NEXUS [12], Geneious
Phylogenetic tree	Phylip (`*.phy`), FASTA, NEXUS [12], MEGA3 [11], Geneious
PDF document	PDF, Geneious

2.2 Acquiring data from public databases

Geneious is able to communicate with a number of public databases hosted by the National Centre for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory (EMBL). You can access these databases through the web at `http://www.ncbi.nlm.nih.gov` and `http://www.ebi.ac.uk/embl/` respectively. Both are well known and widely used storehouses of molecular biology data.

2.2.1 NCBI (Entrez) and EMBL databases

NCBI was established in 1988 as a public resource for information on molecular biology. Geneious allows you to directly download information from seven important NCBI databases, NCBI

BLAST services and one EMBL database (Table 2.1).

Table 2.1: NCBI and EMBL databases accessible via Geneious

Database	Center	Coverage
Genome	NCBI	Whole genome sequences
Nucleotide	NCBI	DNA sequences
PopSet	NCBI	sets of DNA sequences from population studies
Protein	NCBI	Protein sequences
Structure	NCBI	3D structural data
PubMed	NCBI	Biomedical literature citations and abstracts
Taxonomy	NCBI	Names and taxonomy of organisms
Uniprot	EMBL	Protein sequences

This section gives a brief explanation of the key databases that can be searched using Geneious. There is also a bewildering array of acronyms. If you are not already a bioinformatician, please check them before you continue.

Table 2.2: Acronyms

Acronym	Full name
DDBJ	DNA Data Bank of Japan
EMBL	European Molecular Biology Laboratory
cDNA	Complementary DNA
GSS	Genomic survey sequences
PDB	Protein Data Bank
EST	Expressed Sequence Tags
UniProt	the Universal Protein Resource
RefSeq	Reference Sequence
PIR	Protein Information Resource
PRF	Protein Research Foundation
STS	Sequence Tagged Site
HTGS	High Throughput Genomic Sequence

The Entrez Genome database. This provides views of a variety of genomes, complete chromosomes, sequence maps with contigs (contiguous sequences), and integrated genetic and physical maps.

The Entrez Nucleotide database. This database in GenBank contains 3 separate components that are also searchable databases: "EST", "GSS" and "CoreNucleotide". The core nucleotide database brings together information from three other databases: GenBank, EMBI, and DDBJ. These are

part of the International collaboration of Sequence Databases. This database also contains RefSeq records, which are NCBI-curated, non-redundant sets of sequences.

The Entrez Popset database. This database contains sets of aligned sequences that are the result of population, phylogenetic, or mutation studies. These alignments usually describe evolution and population variation. The PopSet database contains both nucleotide and protein sequence data, and can be used to analyze the evolutionary relatedness of a population.

The Entrez Protein database. This database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ as well as protein sequences submitted to the Protein Information Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB) (sequences from solved structures).

The Entrez Structure database. This is NCBI's structure database and is also called MMDB (Molecular Modeling Database). It contains three-dimensional, biomolecular, experimentally or programmatically determined structures obtained from the Protein Data Bank.

The PubMed database. This is a service of the U.S. National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals. This archive of biomedical articles dates back to the 1950s. PubMed includes links to full text articles and other related resources, with the exception of those journals that need licenses to access their most recent issues.

Entrez Taxonomy. This database contains the names of all organisms that are represented in the NCBI genetic database. Each organism must be represented by at least one nucleotide or protein sequence.

UniProt. This database is a comprehensive catalogue of protein data. It includes protein sequences and functions from Swiss-Prot, TrEMBL, and PIR. It has three main components, each optimized for a particular purpose.

The scope and depth of these databases make them critical information sources for molecular biologists and bioinformaticians alike. However, a library is only as good as its librarian. Geneious is your librarian, allowing you to search for, filter and store, only the data that you care about.

2.2.2 Accessing NCBI BLAST through Geneious

BLAST [1] stands for Basic Local Alignment Search Tool. It allows you to query the NCBI sequence databases with a sequence in order to find entries in the public database that contain similar sequences. When "BLAST-ing", you are able to specify either nucleotide or protein sequences and nucleotide sequences can be either DNA or RNA sequences. The result of a BLAST query is a table of "hits". Each hit refers to a GenBank accession number and the gene or protein name of the sequence. Each hit also has a "Bit-score" which provides information about how similar the hit is to the query sequence. The bigger the bit score, the better the match.

Finally there is also an "E-value" or "Expect value", which represents the number of hits with at least this score that you would expect purely by chance, given the size of the database and query sequence. The lower the E-value, the more likely that the hit is real.

Geneious is able to run NCBI BLAST on seven different databases. Some of these databases are non-redundant in order to reduce duplicate hits. You can submit either a raw sequence or Genbank accession number into NCBI BLAST and receive a summary of results for each hit. This summary contains the bit-score, e-value, identity, and the stretch of the query sequence and hit sequence that match. The databases that can be searched are:

Table 2.3: Nucleotide sequence searches in the BLAST databases

Database	Nucleotide searches
env_nt	This contains DNA sequences from the environment, i.e all organisms put together
month	All new/updated GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
nr	All non-redundant GenBank+EMBL+DDBJ+PDB sequences(no EST, STS, GSS or HTGS sequences)
pat	Nucleotide sequences derived from the Patent division of GenBank
PDB	Sequences derived from the 3D-structures of proteins from PDB
RefSeq	NCBI-curated, non-redundant sets of sequences.
SwissProt	-

Table 2.4: Protein sequence searches in the BLAST databases

Database	Protein searches
env_nr	Translations of sequences in env_nt
month	All new/updated GenBank coding region (CDS) translations +PDB+SwissProt+PIR released in last 30 days
nr	All non-redundant GenBank coding region (CDS) translations+PDB+SwissProt+PIR+PRF
pat	Protein sequences derived from the Patent division of GenBank
PDB	Sequences derived from 3D structure Brookhaven PDB
RefSeq	-
SwissProt	Curated protein sequences information from EMBL

Geneious can perform five different kinds of BLAST search:

- **blastp**: Compares an amino acid query sequence against a protein sequence database.
- **blastn**: Compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx**: Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.

- **tblastn**: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **tblastx**: Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.

2.3 Searching databases and local documents using Geneious

Although the above databases are accessible from the NCBI web page, Geneious allows you to comfortably search them from your desktop.

Once you have search results in the Document Table, they can be filtered further with the Filter Box in the Toolbar. Furthermore once you have started to store your sequences and publications as local documents in Geneious you can also search these local folders using the same easy search interface by clicking on the toolbar search icon which opens the a search dialog.

2.3.1 Common search features

The following search features are available under both "Simple Search" and "Advanced Search" options available by clicking on "More Options".

Matching exact words

When you are looking for exact matches to certain words, it is important to put them in quotes "" to reduce the number of incorrect matches. This minimizes the number of irrelevant downloads.

Important: You must use quotation marks ("") if "!", "@", "$", and blank spaces (" ") are part of your search criteria. No quotation marks lead to unreliable results.

Wild card searches

When you are looking for all matches to a partial word, use the asterisk (*). For example, typing "oxi*" would return matches such as oxidase, oxidation, oxido-reductase, and oxide. This is useful for performing generic searches. You can also place the asterisk (*) in the middle of the word.

Please note. Wild card searches only work on searches in the local folders in Geneious.

2.3.2 Search panel

Searching has been designed to be as user-friendly as possible. To search select the ”Search” button from the toolbar to activate search. For non-local folders search will be on by default and cannot be closed. This applies to NCBI and EMBL databases. For local folders search is off by default.

When search is first activated the document table will be emptied to indicate no results have been found. To return to browsing click the ”Search” button again or press the Escape key while the cursor is in the search text field.

To initiate a search enter the desired search term(s) in the text field and press enter or click the adjacent “Search” button. Once a search starts the results will appear in the document table as they are found. The “Search” button changes to a “Cancel” button while a search is in progress and this may be clicked at any time to terminate the search. Feedback on a search progress is presented in the status bar directly below the toolbar.

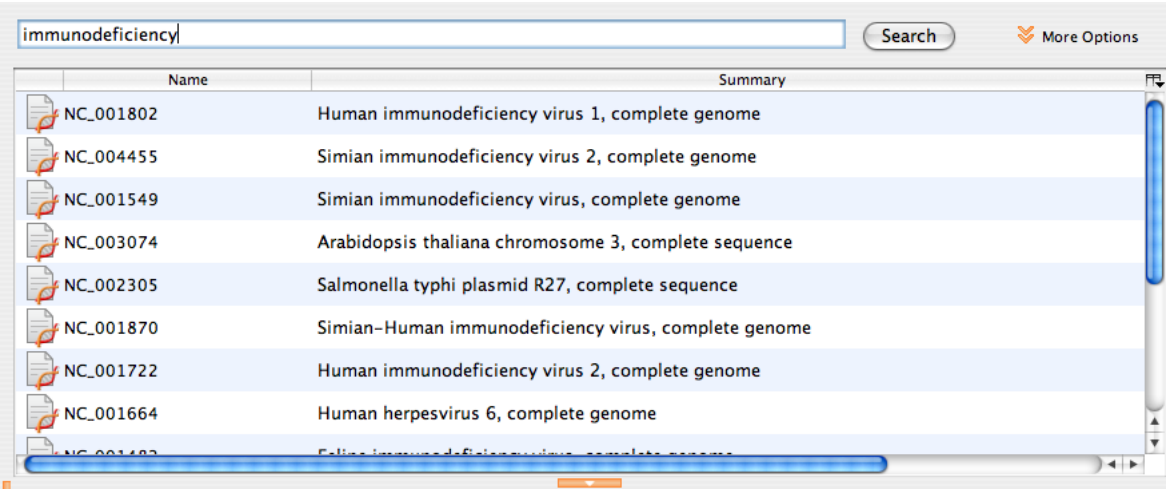


Figure 2.2: The Search tab of the Document Table

2.3.3 Advanced Search options

To access advanced search click the “More Options” button inside the basic search panel. To return to basic search click the “Fewer Options” button. Switching between advanced and basic will not clear the search results table.

This feature provides more search options to select from. Geneious allows you to search with a range of criteria; however, these depend on the database being searched. All the fields in the NCBI public databases can be searched in any combination. Each database has a specific

list of fields and it is important to familiarize yourself with these fields to make full use of the Advanced Search. The fields available for a search can be found in the left-most drop-down box after enabling the advanced search options.

Note. When searching the Genome database, the documents returned are only summaries. To download the whole genome, select the summary(s) of the genome(s) you would like to download and the click the "Download" button inside the document view or just above it. There are also "Download" items in the File menu and in the popup menu when document summary is right-clicked. The size of these files is not displayed in the Documents Table. Be aware that whole genomes can be very large and can take a long time to download. You can cancel the download of document summaries by selecting "Cancel Downloads" from any of the locations mentioned above.

Advanced Search also provides you with a number of options for restricting the search on a field depending on the field you are searching against. For example, if you are using numbers to search for "Sequence length" or "No. of nodes" you can further restrict your search with the second drop-down box:

- "is greater than" ($>$)
- "is less than" ($<$)
- "is greater than or equal to" ($\geq$)
- "is less than or equal to" ($\leq$)

Likewise if you are searching on the "Creation Date" search field you have the following options

- "is before or on"
- "is after or on"
- "is between"

When searching your local folders you have the option of searching by "Document type". The second drop-down list provides the options "is" and "is not". The third drop-down lists the various types of documents that can be stored in Geneious such as "3D-Structure", "Nucleotide sequence", and "PDF" (see Figure 2.3).

And/Or searches

The advanced options lets you search using multiple criteria. By clicking the "+" button on right of the search term you can add another search criteria. You can remove search criteria by

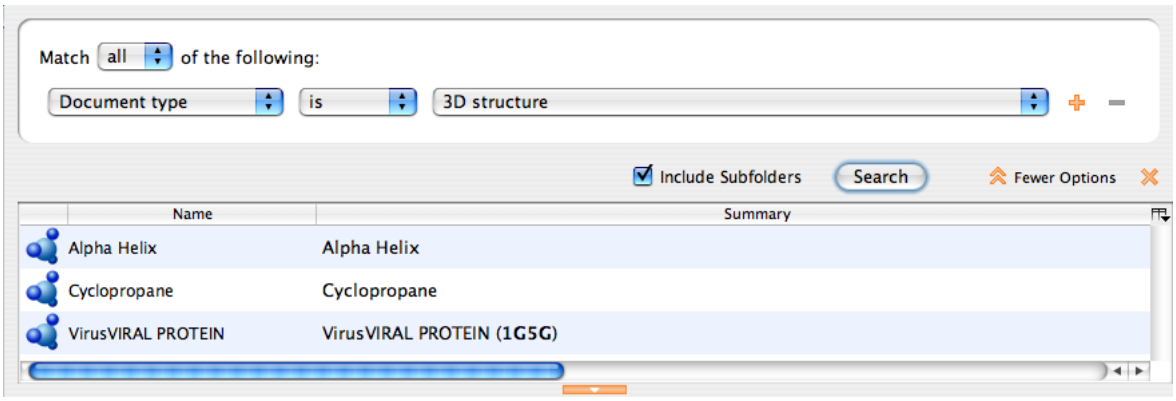


Figure 2.3: Document type search options

clicking on the appropriate "-" button. The "Match all/any of the following" option at the top of the search terms determines how these criteria are combined:

Match "Any" requires a match of one or more of your search criteria. This is a broad search and results in more matches.

Match "All" requires a match all of your search criteria. This is a narrow search and results in fewer matches.

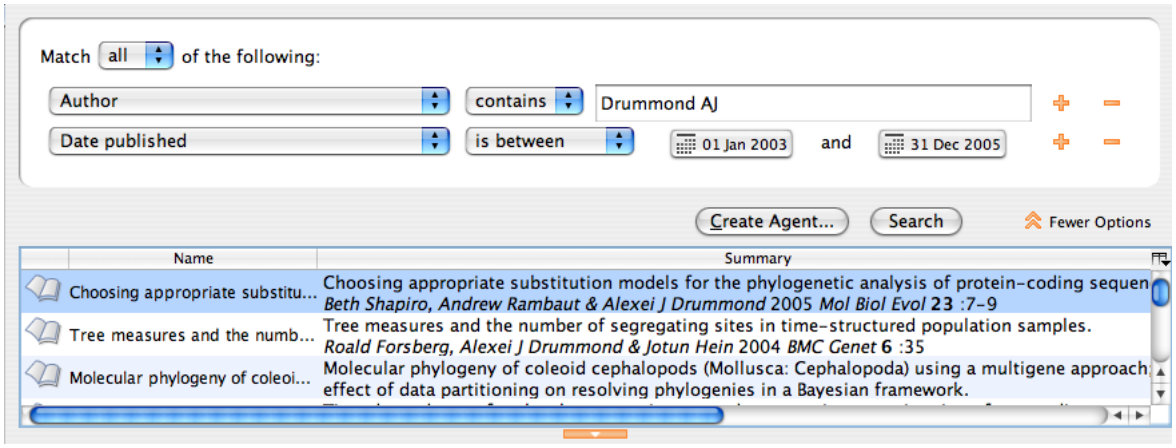

Match all of the following:
Author
contains
Drummond AJ
Date published
is between
01 Jan 2003
and
31 Dec 2005
Create Agent...
Search
Fewer Options
Name
Summary
Choosing appropriate substitu...
Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequen
Beth Shapiro, Andrew Rambaut & Alexei J Drummond 2005 Mol Biol Evol 23 :7-9
Tree measures and the numb...
Tree measures and the number of segregating sites in time-structured population samples.
Roald Forsberg, Alexei J Drummond & Jotun Hein 2004 BMC Genet 6 :35
Molecular phylogeny of coleoi...
Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach
effect of data partitioning on resolving phylogenies in a Bayesian framework.

Figure 2.4: Advanced Search

2.3.4 Autocompletion of search words

Geneious remembers previously searched keywords and offers an auto-complete option. This works in a similar way to Google or predictive text on your mobile phone. If you click within the search field, a drop-down box will appear showing previously used options.

2.4 Agents

Geneious offers a simple way for you to continuously receive the latest information on genomes, sequences, and protein structures. This feature is called an agent. Each agent is a user-defined, automated search. You can instruct an agent to search any Geneious accessible database at regular intervals (eg. weekly) including your contacts on Collaboration. This simple but powerful feature ensures that you never miss that critical article or DNA sequence. To manage agents click on the agent icon in the toolbar. An agent has to be set up before it can be used.

2.4.1 Creating agents

To set up an Agent click the Agents icon and the create button. You now need to specify a set of search criteria in the exact same way as you do for search, the database to search, search frequency and the folder you wish the agent to deliver its results to.

The search frequency may be specified in minutes, hours, days or weeks. You can only use whole numbers.

Selecting "Only get documents created after today" will cause the agent to check what documents are currently available when the agent is created. Then when the agent searches it will only get documents that are new since it was created. e.g. If you have already read all publications by a particular author and you want the agent to only get publications released in the future.

Alternatively you can click the "Create Agent..." button which is available in some advanced search panels. This will use the advanced search options you have entered to create the agent.

The easiest way to organize your search results is to create a new folder and name it appropriately. You can do that by navigating to the parent folder in the "Deliver to" box and click "New Folder", or by creating a new folder beforehand,

1. Right click (for Windows/Unix/Macs with a 2-button mouse) on the "Sample Documents" or "Local" folders. This brings up a popup menu with a "New Folder..." option.
2. Create a new folder and name it according to the contents of the search. (For example, type "CytB" if searching for cytochrome b complex.)

3. Once created, select the new folder. You can now select the "Create" or "Create and Run". The agent will then be added to the list in the agent dialog and it will perform its first search if you clicked "Create and Run". Otherwise it will wait until its next scheduled search.

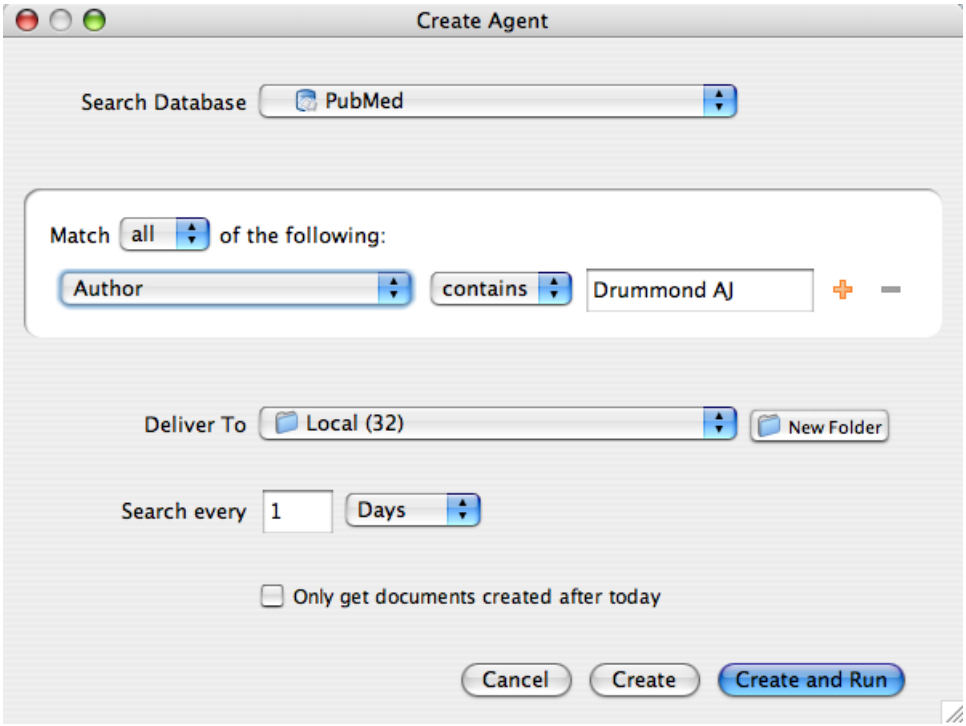


Figure 2.5: The Create Agent Dialog

2.4.2 Checking agents

Once you have created one or more agents, Geneious allows you to quickly view their status in the agents window which is accessible from the toolbar. Your agents' details are presented in several columns: *Enable*, *Action*, *Status* and *Deliver To*.

Enable This column contains a check box showing whether the agent is enabled. *Action*. This summarizes the user-defined search criteria. It contains:

1. Details of the database accessed. For example, Nucleotide and Genome under NCBI.
2. The search type the Agent performed, i.e. "keyword".
3. The words the user entered in the search field for the Agent to match against.

Status. This indicates what the Agent is currently doing. The status will be one of the following:

- "Next search in x time" eg. 18 hours. The agent is waiting until its next scheduled search and it will search when this time is reached.
- "Searching." These are shown in bold. The agent is currently searching.
- "Disabled." The agent will not perform any searches.
- "Service unavailable." The agent cannot find the database it is scheduled to search. This will happen if the database plugin has been uninstalled or if for example the Collaboration contact is offline currently.
- "No search scheduled" The agent is enabled but doesn't have a search scheduled. To correct this click the "Run now" button in the agent dialog to have it search immediately and schedule a new search.

Deliver To. This names the destination folder for the downloaded documents. This is usually your Local Documents or one of your local folders.

Note. If you close Geneious while an agent is running, it will stop in mid-search. It will resume searching when Geneious is restarted. Also, all downloaded files are stored in the destination folder and are marked "unread" until viewed for the first time.

2.4.3 Manipulating an agent

Once an agent has been set up, it can be disabled, enabled, edited, deleted and run. All these options are available from within the Agents dialog.

- *Enable or disable* an agent by clicking the check box in the Enable column.
- "Run Now" Cause the agent to search immediately
- "Cancel" If the agent is currently searching this can be clicked to stop the search.
- "Edit" Click this to change an agent's database, search criteria, destination or search interval.
- "Delete" Delete the agent permanently. Any documents retrieved by the agent will remain in your local documents.

Table 2.5: Geneious document types

Document type	Geneious Icon
Nucleotide sequence	
Protein sequence	
Phylogenetic tree	
3D structure	
Sequence alignment	
Chromatogram	
Journal articles	
PDF	
Other documents	

2.5 Storing data in Geneious

Geneious can be used to store your documents locally. Under the ”Local” folder in the Services Panel you are able to create sub-folders to organize and store a variety of document types (2.5).

This is also where you can set up special folders to receive documents that are downloaded by a Geneious agent. To create a new folder in Geneious, select the “Local” folder or a sub-folder icon in the services panel and right-click (2-button mouse), or Ctrl+click (1-button mouse on Mac OS X). This will pop up a menu. Clicking on “New folder...” opens a dialog that will prompt you to name the folder. The named folder is then created as a sub-folder of the folder that you originally right-clicked on.

Important. Search results will be lost when you exit Geneious unless the downloaded documents have been copied or moved to one of your local folders.

In Geneious you can create new folders, rename existing folders, delete and export folders. All these choices are available by either right-clicking on the folder, clicking on the action menu (Mac OS X), or by holding down the control button and clicking (Mac OS X). Also in Mac OS X, you can also use the plus (+) and minus (-) buttons located at the bottom of the service panel

to create and delete folders.

2.5.1 Transferring data

It is quick and easy to transfer data to your local folders from either a Geneious database search or from your computer's hard drive. Please check you have already set up your destination folders before continuing.

Moving documents from Geneious searches to your Local folders

There are a number of ways to do this.

Drag and drop. This is quickest and easiest. Select the documents that you want to move. Then, while holding the mouse button down, drag them over to the desired folder and release. If you dragged documents from one local folder to another, this action will move the documents – so that a copy of the document is not left in the original location. In external databases such as NCBI the documents will be copied, leaving one in its original location.

Drag and copy. While dragging a document over to your folder, hold the Ctrl key (Alt key on Maje's) down. This places a copy of the document in the target folder while leaving a copy in the original location. This is useful if you want copies in different folders.

The Edit menu. Select the document and then open the Edit menu on the menu bar. Click on "Cut" (Ctrl+X/ Command+X), or "Copy" (Ctrl+C/Command+C). Select the destination folder and "Paste" (Ctrl+V/Command+V) the document into it.

2.5.2 Searching your Local folders

The "Services Panel" allows you to browse your Local folder hierarchy. Next to each folder name in the hierarchy is the number of documents it contains in brackets. When the Local folder or a sub-folder is collapsed (minimized), the brackets next to the folder shows how many files are contained in that folder as well as all of its sub-folders. In addition, if some of the documents in a folder are unread, the number of unread documents will also appear in the brackets.

You can search the Local folder (and sub-folders) the same way you search the public databases by clicking on the "Search" icon. If you have defined a new type of note in Geneious, and a Note has been added, it will also be added to the "Advanced Search" criteria. Look at an example of a new Note type called "Protein size", which takes a text value for the protein in kDa (kiloDaltons) (see Figure 2.6).

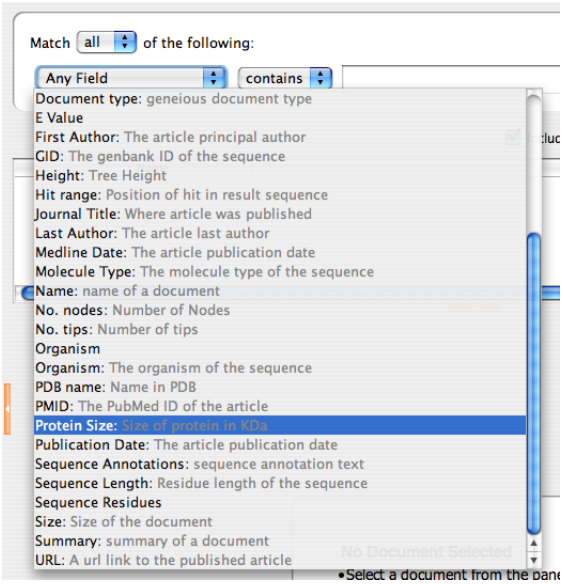

Match all of the following:
Any Field
contains
Document type: geneious document type
E Value
First Author: The article principal author
GID: The genbank ID of the sequence
Height: Tree Height
Hit range: Position of hit in result sequence
Journal Title: Where article was published
Last Author: The article last author
Medline Date: The article publication date
Molecule Type: The molecule type of the sequence
Name: name of a document
No. nodes: Number of Nodes
No. tips: Number of tips
Organism
Organism: The organism of the sequence
PDB name: Name in PDB
PMID: The PubMed ID of the article
Protein Size: Size of protein in KDa
Publication Date: The article publication date
Sequence Annotations: sequence annotation text
Sequence Length: Residue length of the sequence
Sequence Residues
Size: Size of the document
Summary: summary of a document
URL: A url link to the published article
No Document Selected

Figure 2.6: Searching the local documents on a user-defined field

Similarity ("BLAST-like") searching

It is possible to search your local documents not only for text occurrences but by similarity to sequence fragments. Click the small arrow at the bottom of the large T to the left of the search dialog, select "Nucleotide similarity search" or "Protein similarity search" and enter the sequence text. Geneious will try to guess the type of search based on the text, so that simply entering or pasting a sequence fragment may change the search type automatically.

The search locates documents containing a similar string of residues, and orders them in decreasing order of similarity to the string. The ordering is based on calculating an e-value for each match. You can read more about the e-value in subsection 2.2.2.

For the search to be successful, you need to specify a minimum of 11 nucleotides and 3 amino acids. Note that search times depend on the number and size of your sequence documents, and so may take a long time to complete.

2.5.3 Checking and changing the location of your Local folders

To check where your Local folders are being stored on your hard drive, open the Tools menu in the Menu Bar. Click "Tools" → "Preferences" → "General". Your documents are stored at the location specified by the "Data Storage Location" field (see Figure 2.7). You can change this location by clicking the "Browse" button and selecting a new location. Geneious will remember this new location when you exit.

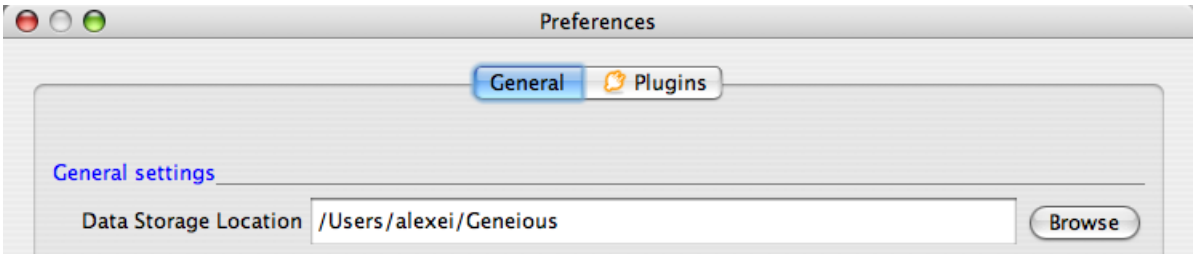


Figure 2.7: Setting the location of your local documents

2.6 Panels in Geneious

This section provides more information on each of the panels in Geneious (Figure 2.8).

2.6.1 The Services Panel

The Services Panel shows a tree that concisely displays Geneious functions and your stored documents. The plus (+) symbol indicates that a folder contains sub-folders. A minus (-) indicates that the folder has been expanded completely and has no sub-folders. Click these symbols to expand or contract folders.

Geneious Service Panel allows you to access:

- Your Local Documents.
- NCBI databases - BLAST [1], Genome, Nucleotide, PopSet, Protein, Pubmed, Structure and Taxonomy.
- An EMBL database - Uniprot.
- Your contacts' Geneious databases.

You can view options for any selected service with the right mouse button, or by clicking the Options button at the bottom of the Service Panel in Mac OS X.

2.6.2 The Documents Table

The Document Table displays your search results or your stored documents. While search results usually contain documents of a single type, a local folder may contain any mixture of documents, whether they are sequences, publications or other types. If you cannot see all of the columns in the document table you may want to close the help panel to make more room.

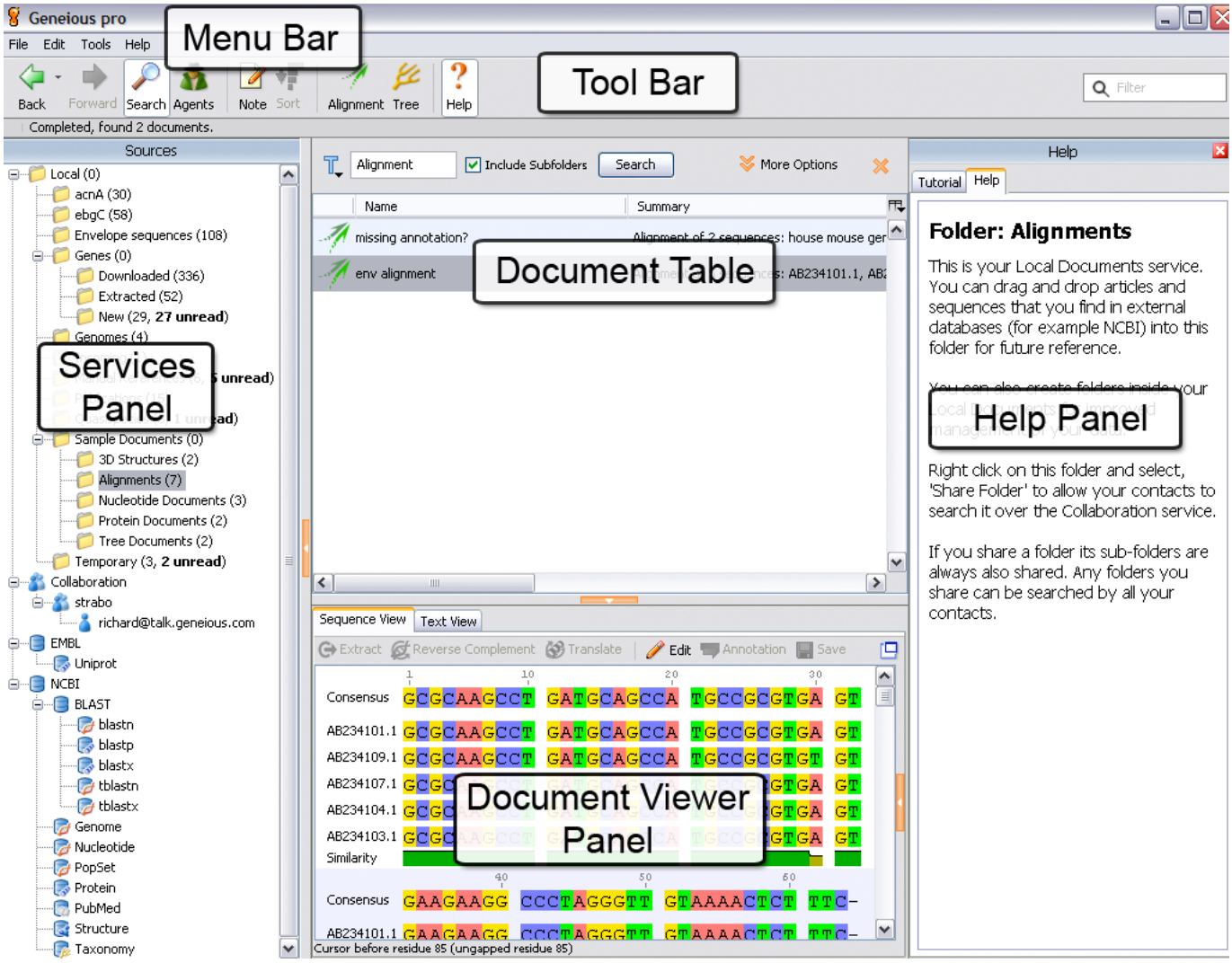


Figure 2.8: Geneious main window

This information is presented in table form (Figure 2.9).

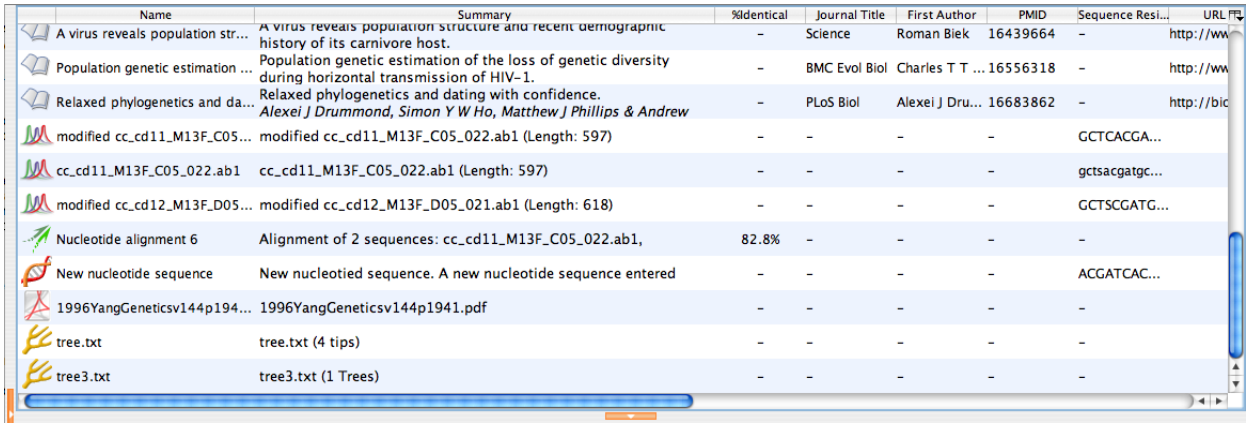


Figure 2.9: The document table, when browsing the local folders

Selecting a document in the Document Table will display its details in the Document View Panel. Selecting multiple documents will show a view of all the selected documents if they are of similar types. eg. Selecting two sequences will show both of them side-by-side in the sequence view. There are several ways to select multiple documents.

- Hold Ctrl (Command on Mac OS) and click to add the document to the current selection.
- Hold Shift and click to add the document and all documents between it and your previous selection.
- On windows the right mouse button can be clicked and held while moving the mouse to easily select a block of documents. The popup menu will appear once the mouse is released so the newly selected documents can quickly be manipulated.

Double-clicking a document in the Document Table displays the same view in a separate window.

To view the actions available for any particular document or group of documents, right-click on a selection of them (Ctrl+Click on Mac OS X). These options vary depending on the type of document.

The Document Table has some useful features.

Editing. Values can be typed into the columns of the table. This is a useful way of editing the information in a document. To edit a particular value, first click on the document and then click on the column which you want to edit. Enter the appropriate new information and press enter. Certain columns cannot be edited however, eg. the NCBI accession number.

Copying. Column values can be copied. This is a quick method of extracting searchable information such as an accession number. To copy a value, right-click on it, and choose the "Copy `name`" option, where `name` is the column name.

Sorting. All columns can be alphabetically, numerically or chronologically sorted, depending on the data type. To sort by a given column click on its header. If you have different types of documents in the same folder, click on the "Icon" column to sort then according to their type.

Rearranging. You can reorder the columns to suit. Click on the column header and drag it to the desired horizontal position.

Choose columns to show The visible columns can be changed by right-clicking (Ctrl-Click on MacOS) on any column header or click the small header button in the top right corner of the table. This gives a popup menu with a list of all the available columns. Clicking on a column will show/hide it. Your preference is remembered so if you hide a column it will remain hidden in all areas of the program until you show it again.

Note. If a Note is added to a document (refer to the section on Adding Notes for more information), a Note column is added to the end of the existing Document table. Also, when accessing BLAST [1] in Geneious, the Document Table has additional columns related to the BLAST search.

2.6.3 The Document Viewer Panel

The Document Viewer Panel shows the contents of any document clicked on in the Document Table. To view large documents, it is sometimes better to double click on them. This opens a view in a new window. In the document viewer panel there are two tabs that are common to most types of documents: "Text view" and "Notes". "Text view" shows the document's information in text format. The exception to this rule occurs with PDF documents where the user needs to either click the "View Document" button or double-click to view it. The "Notes" tab only appears if the user has added notes to the document.

Some document types such as sequences, trees and structures have an options panel occupying the right of the document viewer. The options in the options panel have an arrow which can be used to expand or hide a group of related options.

See the next section on document viewers for more information about operating the various viewers in Geneious.

2.6.4 The Help Panel

The Help Panel has a "Help" tab and a "Tutorial" tab. The Help tab provides you information about the service you are currently using. The Tutorial is aimed at first-time users of Geneious and has been included to provide a feel for how Geneious works. It is highly recommended

that you work through the tutorial if you haven't used Geneious before.

2.6.5 The Toolbar

The toolbar contains several icons that provide shortcuts to common functions in Geneious. You can alter the contents of the toolbar to suit your own needs. The icons can be displayed small or large, and with or without their labels. The Preferences and Help icons are always available. The "Back" and "Forward" options help you move between previous views in Geneious and are analogous to the back and forward buttons in a web browser. The ▽ option shows a list of previous views. The other features that can be accessed form the toolbar are described in chapter 3.

The toolbar can be customized by right-clicking (Ctrl-Click on MacOS) on it. This gives a popup menu with the following options:

- "Show Labels" Turn the text labels on or off.
- "Large Icons" Switch between large and small icons.
- A list of all available toolbar buttons. Selecting/deselecting buttons will show/hide the buttons in the toolbar.

Note. Below the Toolbar, there is a grey status bar. This bar displays the status of the currently selected service. For example, when you are running a search, it displays the number of matches, and the time remaining for the search to finish.

2.6.6 The Menu bar

File Menu. This contains some standard "File" menu items including printing and "Exit" on Windows. On *Geneious pro* it contains "New Sequence" which will create a new nucleotide or protein sequence from residues that you can paste or type in. It also contains options to create, rename, delete or share folders and Import/Export options.

Edit Menu. Here you will find common editing functions including "Cut", "Copy", "Paste", "Delete" and "Select All". These are useful when transferring information from within documents to other locations, or exporting them. This menu also contains "Find in Document", "Find Next" and "Find Previous" options. Find can be used to find text or numbers in a selected document. This is useful when looking for annotated regions or a stretch of bases in a sequence. This opens a "Find Dialog". The shortcut to this is Ctrl+F. *Next* finds the next match for the text specified in the "Find" dialog. The shortcut keys are F3 or Ctrl+G. Geneious then allows you to choose another document and continue searching for the same search word. *Prev* finds the previous match. The shortcut keys for this are Ctrl+Shift+G or Shift+F3.

Tools Menu. This contains a list of features such as "Note", "Alignment", "Tree", "Viewers" and "Preferences". Click on "Viewers" to get a list of all the viewers available in Geneious.

Collaboration Menu (pro *only*). This contains actions that can be performed with Collaboration accounts which allow you to share you work with other Geneious users.

Help Menu. This consists of the standard Help options offered by Geneious.

2.7 Document Viewers in Geneious

2.7.1 The Sequence (and alignment) Viewer

The "Sequence view" tab in the Document Viewer panel is available for Nucleotide sequences, Protein sequences, Alignments and some 3D structure documents. The options available are grouped under headings: "Zoom level", "Annotations", "Colors", "Layout", "Zoom options" and "Statistics". The presence of these options varies with the kind of sequence data being viewed.

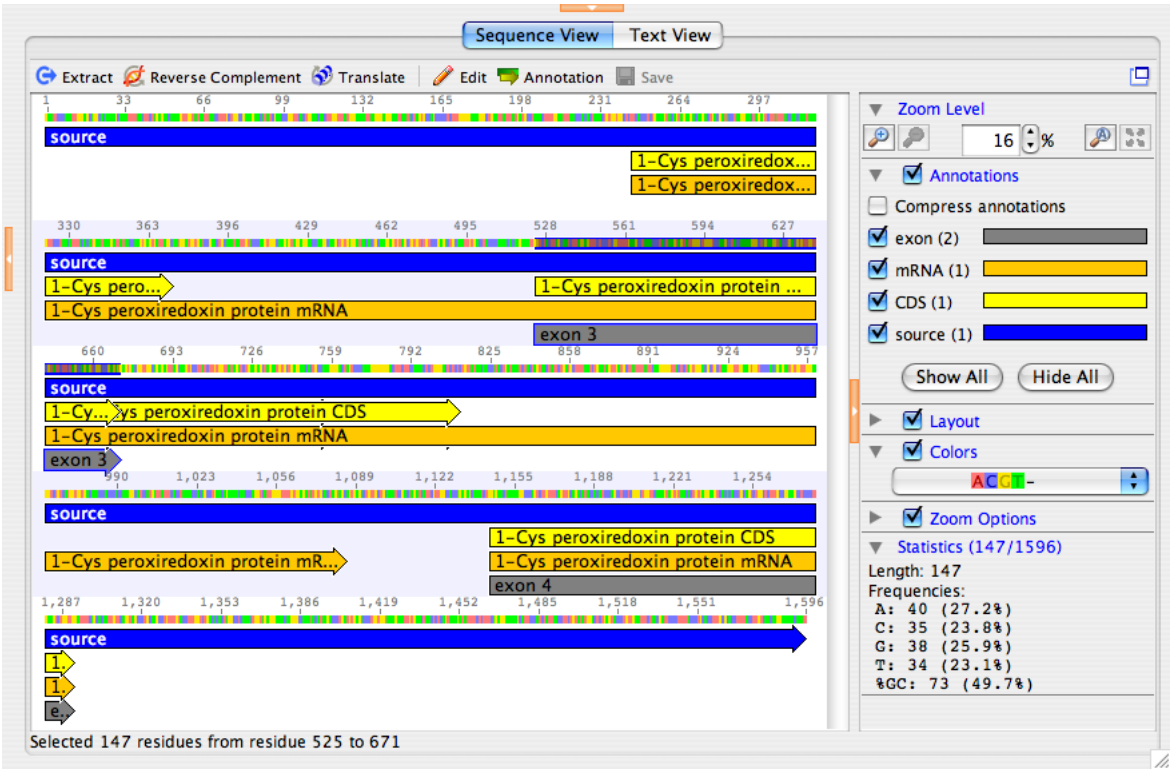


Figure 2.10: A view of an annotated nucleotide sequence in Geneious

Zoom level

The plus and minus buttons increase and decrease the magnification of the sequence by 50%, or by 30% if the magnification is already above 50%.

zooms to 100%. The 100% zoom level allows for comfortable reading of the sequence.

zooms out so as to fit the entire sequence in the available viewing area.

Colors

The colors option controls the coloring of the sequence nucleotides or amino acids. Uncheck the color checkbox to turn off all coloring without viewing further options. Coloring schemes differ depending on the type of sequence. For example, the "Polarity" and "Hydrophobicity" coloring schemes are available only for Protein sequences.

Layout

Layout has various options controlling the layout of the sequence:

- *Show tree*. This toggles the display of the phylogenetic tree when viewing the alignment of a phylogeny document.
- *Show residue positions*. This toggles the display of the residue position number above the sequence residues.
- *Show space every 10 residues*. If you are zoomed in far enough to be able to see individual residues, then an extra white space can be seen every 10 residues when this option is selected.
- *Wrap sequence*. This wraps the sequences in the viewing area. A shortcut is to click the layout check box without expanding it.
- *Wrap on 10-residue boundaries*. This is automatically turned on if the "wrap sequence" option is on and will force the sequence-wrapping to occur in multiples of 10 nucleotides or amino acids.
- *Show sequence and graph names*. Show or hide sequence and graph names inside the sequence viewer panel.

Graphs

This option is visible when viewing protein sequences, chromatogram traces, multiple sequences or sequence alignments. Turn this option on by clicking the Graph checkbox and the graph(s) will be displayed below the sequence(s). A number of graphs are available.

Similarity. This is available for sequence alignments and when more than one nucleotide/protein sequence are displayed. It displays the similarity across all sequences for every position. Green means that the residue at the position is the same across all sequences. Yellow is for less than complete similarity and red refers to very low similarity for the given position (Figure 2.11).

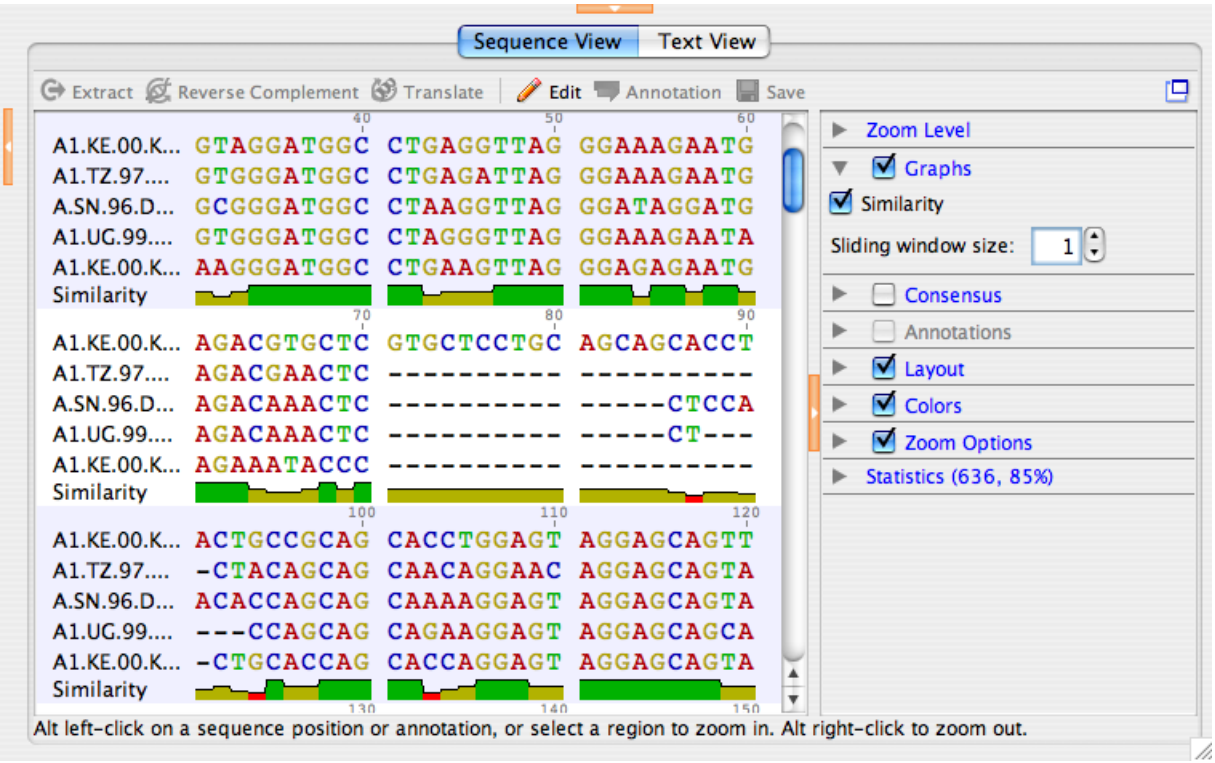


Figure 2.11: The similarity graph for an alignment of nucleotide sequences

Hydrophobicity. This is available with protein sequences. It displays the Hydrophobicity of the residue at every position, or the average Hydrophobicity when there are multiple sequences.

pI. pI stands for Isoelectric point and refers to the pH at which a molecule carries no net electrical charge. The pI plot displays the pI of the protein at every position along the sequence, or the average pI when multiple sequences are being viewed.

Sliding window size. This calculates the value of the graph at each position by averaging across a number of surrounding positions. When the value is 1, no averaging is performed. When the

value is 3, the value of the graph is the average of the residue value at that position and the values on either side.

Chromatogram. This is available with chromatogram traces. It displays the four traces above the sequence, where the peak as detected by the base calling program is at the middle of the base letter. When viewing more than one chromatogram or an alignment made from chromatograms, each chromatogram can be turned on or off individually using the checkbox's below. Note that since the distance between bases as inferred from the trace varies the trace may be either contracted or expanded compared with the raw data.

Show quality. This is available with enabled chromatogram traces. It displays a quality measure (typically Phred quality scores) for each base as assessed by the base calling program. The quality is shown as a shaded bar graph overlaid on top of the chromatogram. Note that those scores represent an estimate of error probability and are on a logarithmic scale - the highest bar represents a one in a million (10^{-6}) probability of calling error while the middle represents a probability of only a one in a thousand (10^{-3}).

Consensus

This option is available when viewing alignments. When checked, the viewer displays the `consensus` sequence with the aligned sequences. The consensus sequences has the same length and shows which residues are conserved (are always the same), and which residues are variable. Basically a consensus is constructed by taking the most frequent residue at each site (position), with some special considerations for ties and gaps.

Ties in a DNA alignment are resolved by taking the appropriate IUPAC nucleotide ambiguity code (i.e R for A or G etc). A tie in a protein alignment is indicated by the general ambiguity symbol X.

no gaps in consensus When this box is checked gaps are not counted and the consensus sequences does not contains any gaps.

Majority consensus When this box is checked only the major character is displayed in the consensus and ambiguity codes only arise in the case of ties. However, when this box is unchecked the relative frequencies of the residues is ignored (i.e. a Strict consensus is performed so that if there is any polymorphism, it is recognized). For example, a site containing 4 G's and one A would have an R as the consensus. This generally implies ignoring gaps since every position normally has at least one non gap residue.

Zoom options

These are a few options that can be turned on or off.

Auto-zoom to selection. If this option is turned on, when you select a range of sequence residues,

the sequence viewer automatically zooms in (or out) so that the selected piece fills the entire viewing area. A shortcut is to select the "zoom options" checkbox without expanding it.

Default zoom level. This options allows the user to specify the initial level of zoom when viewing a sequence.

Please note. Geneious automatically restores the previous zoom level, and over-rides the default settings, when you return to a sequence that you were previously viewing.

Statistics

This displays some statistics about the sequence being viewed. They correspond to the sequence/alignment being viewed or the highlighted part of the sequence/alignment. The length of the sequence or part of the sequence is displayed next to the Statistics option.

Residue frequencies. This section lists the residues for both DNA and amino acid sequences, and also for alignments. It gives the frequency of each nucleotide or amino acid over the entire length of the sequence, including gaps. If there are gaps, then a second percentage frequency is calculated ignoring gap characters. The G+C content for nucleotide sequences is shown as well for easy reference.

The following statistics are available when viewing alignments or multiple sequences,

Pairwise % Similarity The average percent similarity over the alignment. This is computed by looking at all pairs of bases at the same column and scoring a hit (one) when they are identical, divided by the total number of pairs.

Identical sites. The number of sites that are identical across all sequences.

Annotations

Some protein and nucleotide sequences come with annotations and these can be viewed within Geneious sequence viewer. In the presence of annotations, the options panel includes an "Annotations" check box (Figure 2.12). Uncheck the check box to turn off all annotations. Individual annotation types can be turned on or off by using the check boxes next to them.

Compress annotations. This option reduces the vertical height of the annotations on display. This reduces the space occupied by annotations by allowing them to overlap and increases the amount of the sequence displayed on the screen.

Hide all/Show all. These buttons can be used to turn off and on all annotations on the sequence.

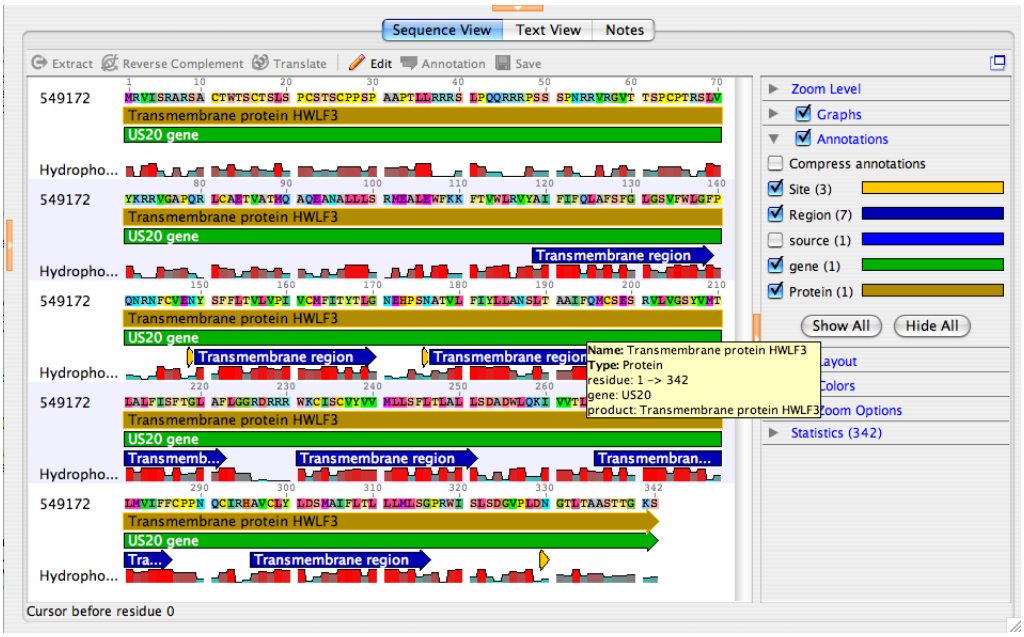


Figure 2.12: The annotations options in the sequence viewer

The sequence viewer toolbar

The top of the sequence viewer panel shows a toolbar containing several actions. Some of them operate on a part of a sequence or alignment. There are several ways to make such a selection.

- *Mouse dragging*. Click and hold down the left mouse button at the start position, and drag to the end position.
- *Select from annotations* When annotations are available, click on any annotation to select the annotated residues.
- *Click on sequence name*. This will select the whole sequence.
- *Select all*. Use the keyboard shortcut Ctrl+A to select everything in the panel.

The available actions are,

Extract Extract the selected part of a sequence or alignment into a new document.

Reverse Complement Reverse sequence direction and replace each base by its complement. This is available only for nucleotide sequences.

Translate. Translate DNA into protein. Clicking on this choice brings up a list of genetic codes that can be used. Choose the appropriate one and click OK. This is available only for nucleotide

sequences.

Edit, Annotations and Save

Editing sequences and alignments

To edit sequence(s) or an alignment click the "Edit" toolbar button. After selecting a residue or a region you can either type in the new contents or use any of the standard editing operation such as Copy (Ctrl+C), Cut (Ctrl-X), Paste (Ctrl-V) and Undo (Ctrl+Z). All operations are under the main "Edit" menu.

Selecting a region enables the "Annotations" button as well, which opens an annotation entry dialog. Enter an annotation name and select a existing type or type a new one. Click on "More Options" to enter additional properties for that annotation. Double click on an existing annotation to edit it or right-click (Ctrl+click on Mac OS X) to display a pop-up menu to delete annotations. You can also copy an annotation from one sequence to another from the pop-up menu.

When editing an alignment it is possible to select a region (which may span several sequences) and drag it to the left or right. Dragging will either move residues over existing gaps or open new gaps when necessary. Dragging a selection consisting entirely of gaps moves the gaps to the new location.

To quickly select a single residue, double-click on it. Triple clicking will select a block of residues within a single sequence. Quadruple clicking selects a block of residues in multiple sequences.

After editing is complete, click "Save" to permanently save the new contents.

The Pop up menu in the sequence viewer

The toolbar actions are available via a pop-up menu as well. Right click (Ctrl+click on Mac OS X) on any sequence, partly highlighted sequence, or annotation to show the various options. The pop-up menu contains the "Copy residues" action (keyboard Ctrl+C) to copy the selected residues to the system clipboard.

Printing a sequence view

To print a sequence view, go to "File" → "Print" and click "OK". The view is printed without the options panel. It is recommended to turn on "Wrap sequence" and deselect "Colors" before printing. Wrapping prints the sequence as seen in the sequence viewer and the font size is chosen to fill the horizontal width of the page.

2.7.2 Dotplot viewer

This is a special viewer that appears when two sequences are chosen. A dotplot compares two sequences to find regions of similarity. Each axis (X and Y) on the plot represents one of the sequences being compared (Figure 2.13). For more information about interpreting dot plots refer to the section on them in chapter 3.

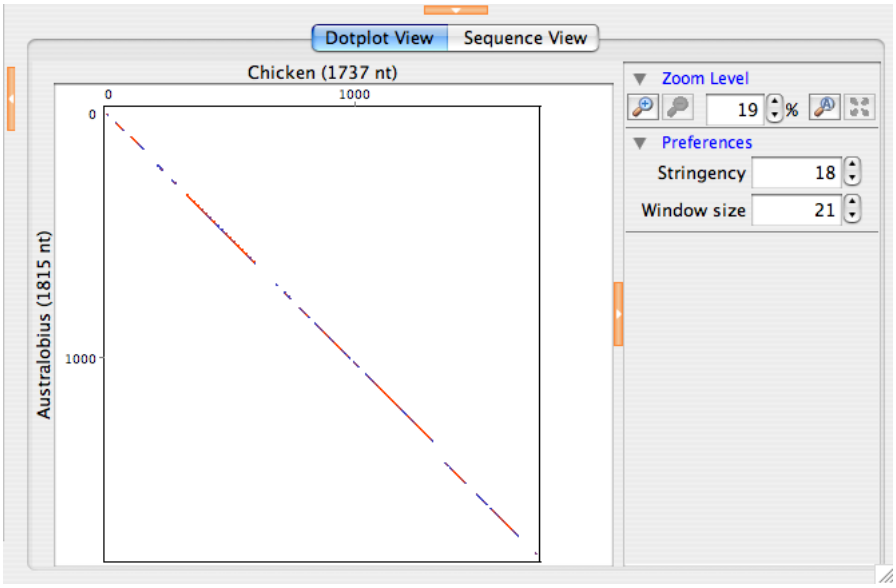


Figure 2.13: A view of dotplot of two sequences in Geneious

2.7.3 3D structure viewer

The 3D structure viewer was introduced in Geneious 1.0 (Figure 2.14). 3D structures can be obtained by searching NCBI's Structure database from within Geneious or by importing 3D structure files (such as PDB files, but other formats are supported as well) from your hard drive. To view the molecule from another angle just click on the molecule and drag the mouse to rotate it. If you would like to change the zoom level of the molecule just hold the shift key and the mouse button, and then drag up and down to zoom out and zoom in. Another way is to use the scroll button on your mouse. Scrolling down enlarges the image and scrolling upwards shrinks it. Under Windows and Linux, you can also move the molecule parallel to the viewing plane while pressing Ctrl+Alt+left mouse button. This is currently not possible under Mac OS.

There are a several viewing options available for 3D structures.

Show atoms. This shows the atoms on the molecule. You can use the "Size" option to increase

or decrease the size of the atoms.

Show bonds. This shows the bonds between the atoms of the molecule.

Show ribbons. This option displays the secondary structure of the molecule in the form of ribbons. This can be viewed only if the necessary information is available.

Show hydrogens. This option displays the hydrogens attached to the molecule. They are shown in white.

Spin. This allows the molecule to spin on a vertical axis for viewing.

Atom symbols. This displays single-letter codes for the atoms adjacent to each atom.

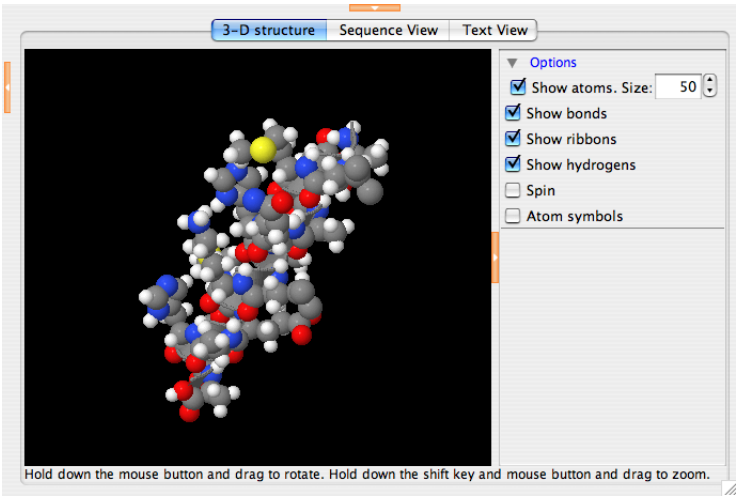


Figure 2.14: A view of a 3D protein structure in Geneious

2.7.4 Tree viewer

The tree viewer provides a graphical view of a phylogenetic tree (Figure 2.15). When viewing a tree a number of other view tabs may be available depending on the information at hand. The "Sequence View" tab will be visible if the tree was built from a sequence alignment using Geneious. The "Text View" shows the tree in text format (Newick). The Notes tab will be present if you have added notes to the tree document.

There are a number of options for the tree viewer.

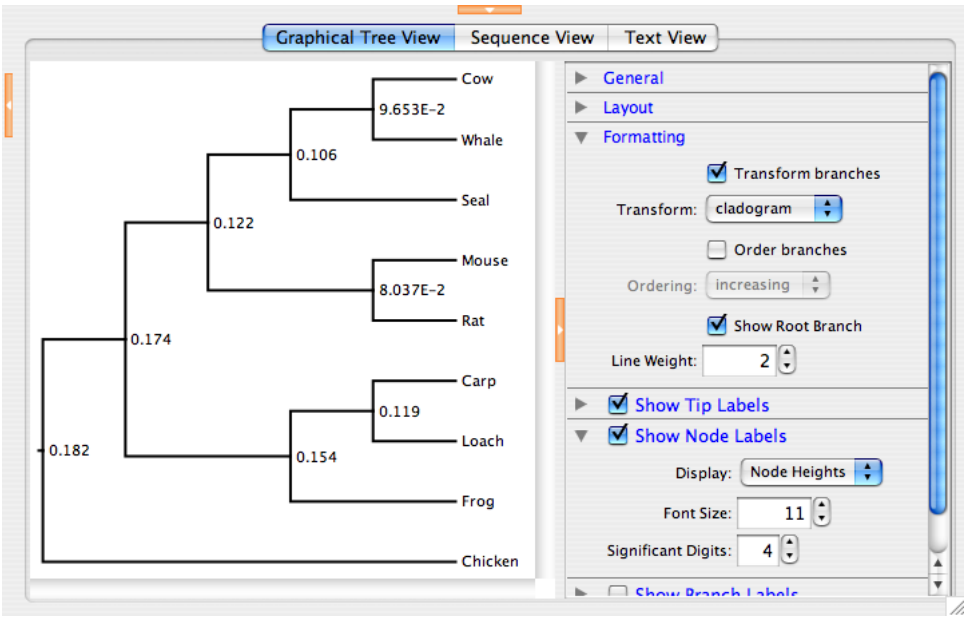


Figure 2.15: A view of a phylogenetic tree in Geneious

General

"General" has 3 buttons showing the different possible tree views: rooted, circular, and unrooted. The "Zoom" slider controls the zoom level of the tree while the "Expansion" slider expands the tree vertically.

Layout

This has two options: root length and curvature. Both "Root Length" and "Curvature" can be increased by dragging the bar towards the right. Root length is not available for unrooted trees. Checking the "Align taxon labels" box will align the taxa labels to provide easier viewer for large trees.

Formatting

There are a range of formatting options.

Transform branches allows the branches to be equal like a cladogram, or proportional. Leaving it unselected leaves the tree in its original form.

Ordering orders branches in increasing or decreasing order of length, but within each clade or

cluster.

Show root branch displays the position of the root of the tree.

Line weight can be increased or decreased to change the thickness of the lines representing the branches.

If you are unfamiliar with tree structures, please refer to Figure 2.16 for the following options.

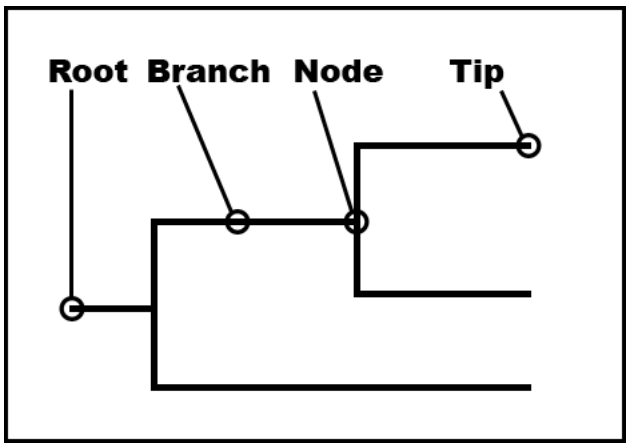


Figure 2.16: Phylogenetic tree terms

Show tip labels. This refers to labels on the tips of the branches of the tree.

Show node labels. This refers to labels on the internal nodes of the tree.

Show branch labels. This refers to the branches of the tree.

Show scale bar. This displays a scale bar at the bottom of the tree view to indicate the length of the branches of the tree. It has three options: "Scale range", "font size" and "line weight".

Depending on the option, "Display" can be branch lengths, taxon names, or node heights. You can use "Font" to change the size of the labels.

2.7.5 The Chromatogram viewer

The Chromatogram viewer provides a graphical view of a the output of a DNA sequencing machine such as Applied Biosystems 3730 DNA analyzer. The raw output of a sequencing machines is known as a `trace`, a graph showing the concentration of each nucleotide against sequence positions. The raw trace processed by a "Base Calling" software which detects peaks in the four traces and assigns the most probable base at more or less even intervals. Base calling may also assign a quality measure for each such call, typically in terms of the expected probability of making an erroneous call.

Sequence Logo. When checked, bases letters are drawn in size proportional to call quality, where larger implies better quality or smaller chance of error. Note that the scale is logarithmic: the largest base represents a one in a million (10^{-6}) or smaller probability of calling error while half of that represents a probability of only a one in a thousand (10^{-3}).

Mark calls. Draw a vertical line showing the exact location of the call made by the base calling software.

Layout. Options controlling layout and view. Those include X and Y axis scaling, size of largest base letter (when Sequence logo is on) and minimum size of base letter (to prevent bases of low quality becoming unreadable).

2.7.6 The PDF document viewer

This viewer is under development. Currently, to view a `.pdf` document either double click on the document in the Documents Table or click on the "View Document" button. This opens the document in an external PDF viewer such as Adobe Acrobat Reader or Preview (Mac OS X). On Linux, you can set an environmental variable named "PDFViewer" to the name of your external PDF viewer. The default viewers on Linux are `kpdf` and `evince`.

2.7.7 The Journal Article Viewer

This viewer provides two tabs: "Text View" and "BibTex". "Text view" displays the journal article details including the abstract. The text contains a link to the original article through Google Scholar below the title and authors (Figure 2.17). BibTex is the standard LaTeX bibliography reference and publication management data format. LaTeX is a common program used to create formatted documents including this one. The information in the BibTex screen can be exported for use in LaTeX documents.

2.8 Preferences

There are two sets of preferences: General and Plugins. You can access the preferences screen in two ways:

1. Shortcut keys: Ctrl+Shift+P (Windows/Linux), Command+Shift+P (Mac OS X)

2. Select the Tools Menu and click Preferences.

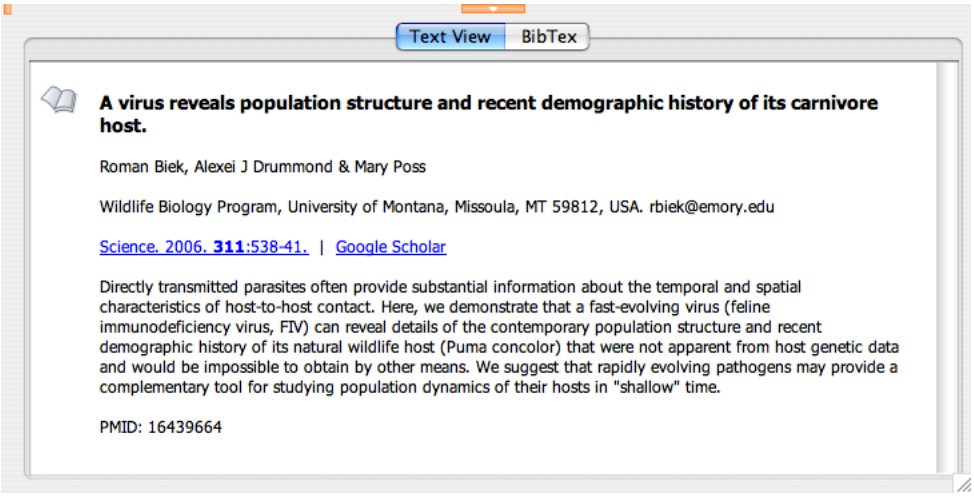

Text View
BibTex
A virus reveals population structure and recent demographic history of its carnivore host.
Roman Biek, Alexei J Drummond & Mary Poss
Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA. rbiek@emory.edu
Science. 2006. 311:538-41. | Google Scholar
Directly transmitted parasites often provide substantial information about the temporal and spatial characteristics of host-to-host contact. Here, we demonstrate that a fast-evolving virus (feline immunodeficiency virus, FIV) can reveal details of the contemporary population structure and recent demographic history of its natural wildlife host (Puma concolor) that were not apparent from host genetic data and would be impossible to obtain by other means. We suggest that rapidly evolving pathogens may provide a complementary tool for studying population dynamics of their hosts in "shallow" time.
PMID: 16439664

Figure 2.17: Viewing bibliographic information in Geneious

2.8.1 General

This contains connection settings, data storage details for your Local documents, Automatic new version checking and “Search History”.

”Check for new version of Geneious” Enable this to have Geneious check for the release of new versions everytime it is started. If a new version has been released Geneious will tell you and give you a link to download it.

”Also check for beta version of Geneious” Enable this to also have Geneious alert you when new beta versions are released. A beta version is a version that is released before the official release for the purposes of testing. It may therefore be less stable than official releases.

Search History. This clears all the previously searched words in Geneious. After this, the auto-completion drop-box will be empty.

Connection settings. These are described in the troubleshooting section of the manual.

2.8.2 Plugins

The “Plugins” tab (Figure 2.18) contains a table of the currently available plugins for Geneious. To enable a plugin, select the checkbox next to it. To disable a plugin, deselect the checkbox next to it.

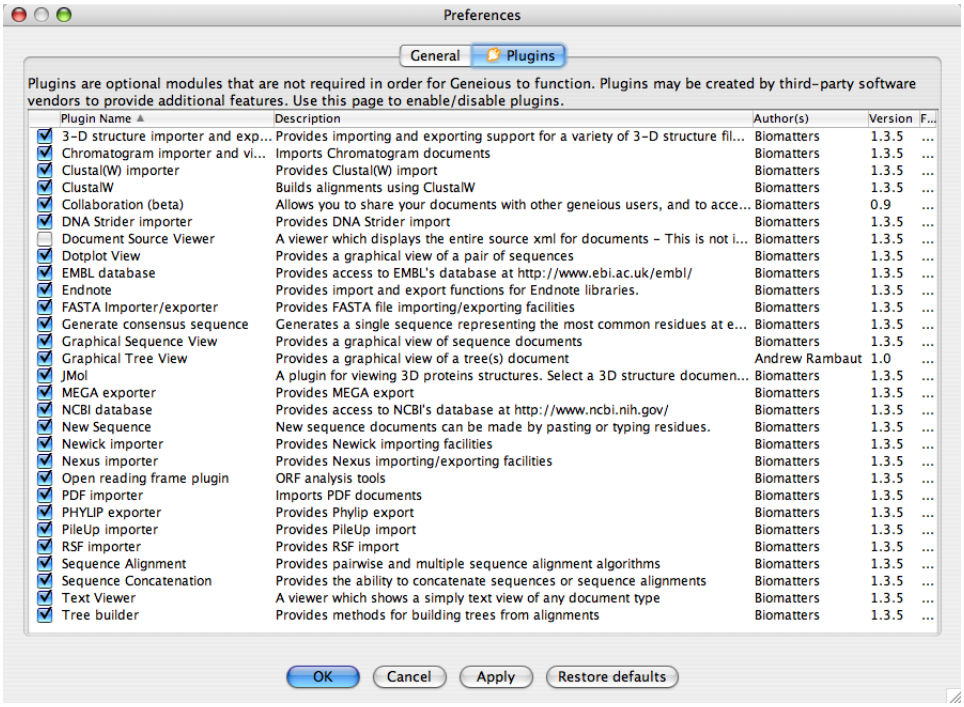

Preferences
General
Plugins
Plugins are optional modules that are not required in order for Geneious to function. Plugins may be created by third-party software vendors to provide additional features. Use this page to enable/disable plugins.
Plugin Name
Description
Author(s)
Version
F...
3-D structure importer and exp... Provides importing and exporting support for a variety of 3-D structure fil... Biomatters 1.3.5 ...
Chromatogram importer and vi... Imports Chromatogram documents Biomatters 1.3.5 ...
Clustal(W) importer Provides Clustal(W) import Biomatters 1.3.5 ...
ClustalW Builds alignments using ClustalW Biomatters 1.3.5 ...
Collaboration (beta) Allows you to share your documents with other geneious users, and to acce... Biomatters 0.9 ...
DNA Strider importer Provides DNA Strider import Biomatters 1.3.5 ...
Document Source Viewer A viewer which displays the entire source xml for documents - This is not i... Biomatters 1.3.5 ...
Dotplot View Provides a graphical view of a pair of sequences Biomatters 1.3.5 ...
EMBL database Provides access to EMBL's database at http://www.ebi.ac.uk/embl/ Biomatters 1.3.5 ...
Endnote Provides import and export functions for Endnote libraries. Biomatters 1.3.5 ...
FASTA Importer/exporter Provides FASTA file importing/exporting facilities Biomatters 1.3.5 ...
Generate consensus sequence Generates a single sequence representing the most common residues at e... Biomatters 1.3.5 ...
Graphical Sequence View Provides a graphical view of sequence documents Biomatters 1.3.5 ...
Graphical Tree View Provides a graphical view of a tree(s) document Andrew Rambaut 1.0 ...
JMol A plugin for viewing 3D proteins structures. Select a 3D structure documen... Biomatters 1.3.5 ...
MEGA exporter Provides MEGA export Biomatters 1.3.5 ...
NCBI database Provides access to NCBI's database at http://www.ncbi.nih.gov/ Biomatters 1.3.5 ...
New Sequence New sequence documents can be made by pasting or typing residues. Biomatters 1.3.5 ...
Newick importer Provides Newick importing facilities Biomatters 1.3.5 ...
Nexus importer Provides Nexus importing/exporting facilities Biomatters 1.3.5 ...
Open reading frame plugin ORF analysis tools Biomatters 1.3.5 ...
PDF importer Imports PDF documents Biomatters 1.3.5 ...
PHYLIP exporter Provides Phylip export Biomatters 1.3.5 ...
PileUp importer Provides PileUp import Biomatters 1.3.5 ...
RSF importer Provides RSF import Biomatters 1.3.5 ...
Sequence Alignment Provides pairwise and multiple sequence alignment algorithms Biomatters 1.3.5 ...
Sequence Concatenation Provides the ability to concatenate sequences or sequence alignments Biomatters 1.3.5 ...
Text Viewer A viewer which shows a simply text view of any document type Biomatters 1.3.5 ...
Tree builder Provides methods for building trees from alignments Biomatters 1.3.5 ...
OK
Cancel
Apply
Restore defaults

Figure 2.18: The plugins preferences in Geneious

Chapter 3

Data Processing

This chapter details the different ways that data can be filtered and analyzed in Geneious.

By the end of this chapter you should be able to:

- Filter data using the tools in Geneious
- Know about the advantages of the "Add Note" function
- Understand the basic principles of bioinformatics
- Perform simple bioinformatics analyses with Geneious

3.1 Data Filtering

The public bioinformatics databases are large and growing rapidly and often contain redundant data and partial matches. This makes it extremely difficult to locate only the most relevant information. Geneious makes this task simple through several user-friendly tools such as Notes, Filtering, Text Matching, and Similarity Searching.

3.1.1 Notes

"Add Note" is a unique feature offered by Geneious. Any Notes that you add can be treated as user-defined fields for use in sorting, searching and filtering your documents.

Where can I add Notes?

The "Add Note" function can be used to add notes to all the data types that Geneious can handle, including molecular sequences, phylogenetic trees and journal articles.

How do I add Notes?

There are 3 ways to open the "Add Note" dialog,

1. *Right-click on a document*. Select a document and right-click (Ctrl+click for Macs) to open an options menu. Scroll down to "Note" and select it.

2. *Through the Toolbar*. Click the Note icon on the Toolbar.

3. *From the "Tools" menu*. Click on the Tools menu on the Toolbar. Scroll down to "Add Note" and click.

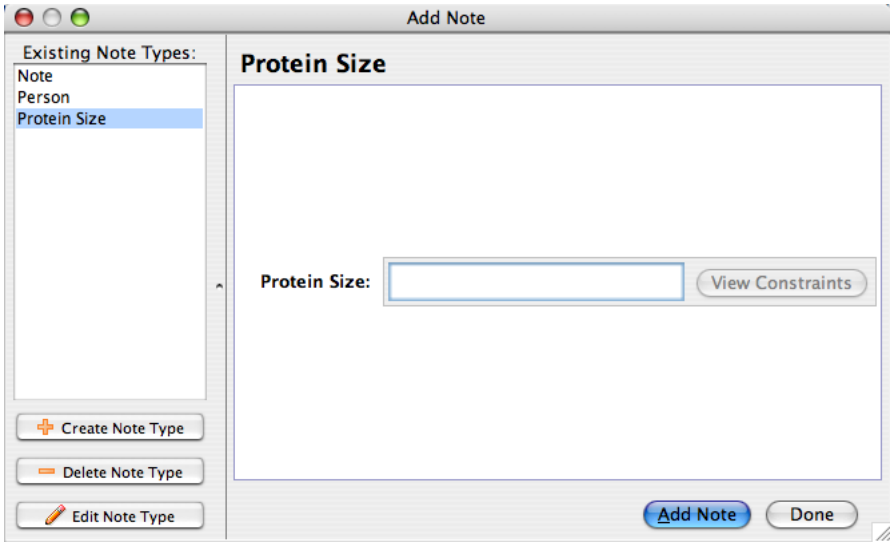


Figure 3.1: The Add Note Menu

The top-left panel lists the existing Notes types. Click on a Note type and add your note. For example, if you choose to add a note of the type "Person", you could add your name to record that the sequence was produced by you.

Creating Note Types

Geneious does not restrict you to certain note types. You can create new note types to annotate your documents.

To create a new note type, select a document and open the Add Note dialog. Click on the Create Note Type button (+) in the left-hand panel of this window. This brings up a window similar to that displayed in 3.1.

Note. The "Note Type Name" and "Note Type Description" fields distinguish your Note type from other user-defined note types. They do not have any constraints. Here are some examples of Note Types.

Name	Description	Data type
Protein size	Size of the protein in kDa	Protein sequences
Tree building method	Method used to build tree UPGMA/Neighbor joining	Phylogenetic trees

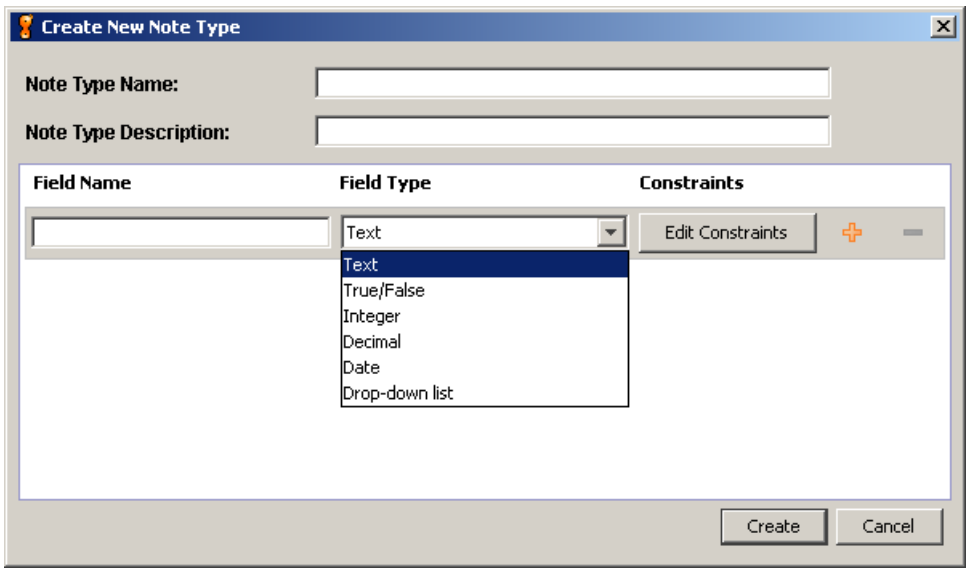


Figure 3.2: Creating new Note Types

You need to decide what other characteristics your Note Type will have.

Field name. This defines what the field will be called. It will be displayed alongside columns such as Description and Creation Date in the Documents Table. You can have more than one Field in a single Note Type.

Field type. This describes the kind of information that the column contains such as Text, Integer, and True/False. The full list of choices in Geneious is shown in 3.2.

Constraints. These are limiting factors on the data and are specific to each field type. For ex-

ample, numbers have numerical constraints – is greater than, is less than, is greater or equal to, and is less or equal to. These can be changed to suit. The constraints for each field can be viewed by clicking the "View Constraints" button next to the field. This will show a pop-up menu with the constraints you have chosen. (3.3)

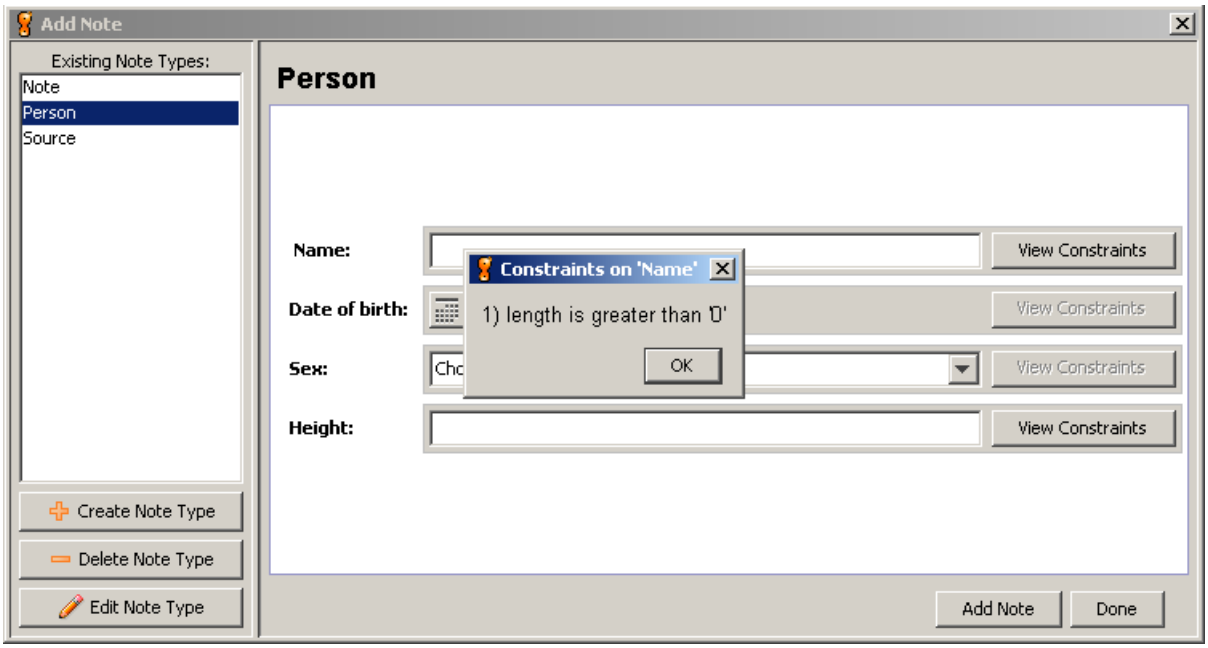


Figure 3.3: Viewing constraints for a field

Using Note Types

The main purpose of Notes is to add user defined information to Geneious documents. However, Notes and Note Types can be searched for and filtered as well. Also, columns can be ordered according to the values of an added Note.

Searching. Once a Note Type is defined and a Note of that type added, it is automatically added to the standard search fields. These are listed under the "Advanced Search" options in the Document Table. From then on, you can use them to search your Local Documents. If you have more than one Field Type for a Note Type, they will both appear as searchable fields in the search criteria.

Filtering. Note values can be used to filter the documents being viewed. To do so, type a value of your Note Type into the "Filter Box" in the right hand side of the Toolbar. Only matching documents will be shown.

Ordering columns. The fields and values of an added Note Type will appear as columns in the

Document Table. These new columns can be used to order the table. Take the example of protein size. A click on the column heading will order the documents in increasing or decreasing order according to their protein size. Clicking the column heading again arranges the documents in the opposite order. An arrow next to the heading indicates if it is in increasing (∧) or decreasing (∨) order.

Viewing Notes

Notes are added into columns defined by their Types and Fields. Notes are also shown as an extra tab in the Document Viewer Panel. Click on the Notes tab to view your notes. Notes can be edited or deleted from here, too. (3.4.)

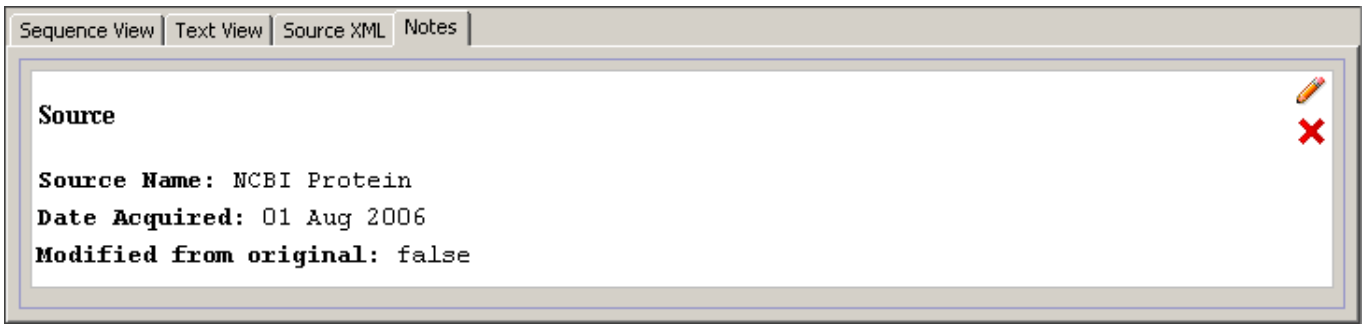


Figure 3.4: Viewing Notes

Editing and Deleting Notes (and Note Types)

Notes can be edited or deleted from the Documents Viewer Panel and from the Add Note screen. Click on the "Edit Note Type" or "Delete Note Type" button (3.5.)

3.1.2 Filtering

The "Filter" is a special feature offered by Geneious. It is in the top right hand corner of the Main Toolbar and is used to filter documents based on any information listed in the Documents Table.

Type in the text you are searching for and Geneious will display all the documents that match this text and hide all other documents in the Document Table. To view all the documents in a folder, clear the Filter box of text or click the button.

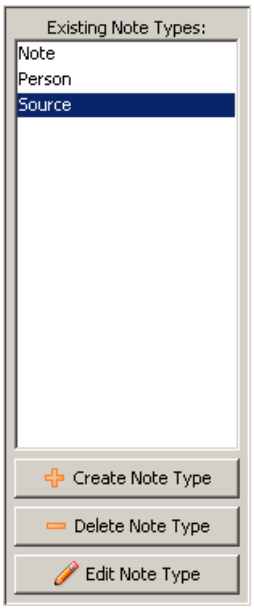


Figure 3.5: Viewing Notes

Filtering on-the-fly

Filtering can be used while searching for documents via public databases, filtering data as it is being downloaded. Type in the appropriate text in the Filter Box and only those documents that match both the original criteria (as specified by the search terms) and the "Filter" text will be displayed. This is an effective way of filtering within your search results.

3.2 Literature

Geneious allows you to search for relevant literature in NCBI's PubMed database. The results of this search are summarized in columns in the Document Table and include the PubMed ID (PMID), first and last authors, URL (if available) and the name of the Journal. When a document is selected, the abstract of the article is displayed in the Document Viewer along with a link to the full text of the document if available, and a link to Google Scholar, both below the author(s) name(s).

Note: If the full text of the article is available for download in PDF format, it can also be stored in Geneious by saving it to your hard drive and then importing it. This will allow full-text searches to be performed on the article.

As well as the abstract and links, Geneious also shows the summary of the journal article in BibTex format in a separate tab of the Document Viewer. This can be imported directly into a

LaTeX document when creating a bibliography. Alternatively, a set of articles in Geneious can be directly exported to an EndNote 8.0 compatible format. This is usually done when creating a bibliography for Microsoft Word documents.

3.3 Sequence data

Basic techniques, such as dotplots and pairwise alignments, can be used to study the relationships between two sequences. However, as the number of sequences increases, methods for determining the evolutionary relationships between them become more complicated.

When analyzing more than two sequences, there are some common steps to determine the ancestral relationships between them. The next four sections outline the basic tools for preliminary sequence analysis: dot plots, sequence alignment and phylogenetic tree building.

3.4 Dotplots

A dotplot compares two sequences against each other and helps identify similar regions [13]. Using this tool, it can be determined whether a similarity between the two sequences is global (present from start to end) or local (present in patches).

The dotplot uses a window of comparison to determine the level of similarity between every pair of sub-sequences (of length window size) in the two sequences being compared.

The dots in a dotplot are determined by two factors: stringency and window size. The stringency is the number of matches required in the given window size for a dot to be plotted and it acts as a threshold that determines the sensitivity of the dot plot. Reducing the required number of matches (for a given window size) will increase the sensitivity of the dotplot, but will also lead to more false positives (regions that match due to chance alone).

If the stringency is set to 3 and the window size to 5, then a pair of 5-base windows that contain 3 or more matching nucleotides will be classified as a match. A dot corresponding to the center of the two windows will then be displayed in the dotplot. Anything less will be classified as a mismatch and no dot will be drawn.

3.4.1 Viewing Dotplots

To view a dotplot in Geneious, select two nucleotide or protein sequences in the Document Table and select Dotplot Viewer in the Document Viewer Panel (see Figure 2.13). For more details on the dotplot viewer and the available options, refer to section 2.7.2.

3.4.2 Interpreting a Dotplot

- Each axis of the plot represents a sequence.
- A long, largely continuous, diagonal indicates that the sequences are related along their entire length.
- Sequences with some limited regions of similarity will display short stretches of diagonal lines.
- Diagonals on either side of the main diagonal indicate repeat regions caused by duplication.
- A random scattering of dots reflects a lack of significant similarity. These dots are caused by short sub-sequences that match by chance alone.

For more information on dotplots, refer to the paper by Maizel & Lenk [13].

3.5 Pairwise sequence alignments

A pairwise sequence alignment is an attempt to determine the regions of homology in two sequences. Over evolutionary time, related DNA sequences diverge through the accumulation of nucleotide substitutions, insertions and deletions. As a result, an alignment is essential to evaluate the degree of similarity between two related sequences. If two nucleotides or amino acids are present in the same column of a pairwise sequence alignment, they are "aligned." This implies that they are homologous, and have evolved from a common ancestral nucleotide/amino acid.

While the aligned nucleotides indicate common ancestry, the regions of a pairwise alignment that contain gaps represent areas where insertions or deletions have occurred in the evolutionary history of one or both of the sequences.

Pairwise alignments can be of two main types: local and global.

A Local Alignment. A local alignment is an alignment of two sub-regions of a pair of sequences [19]. This type of alignment is appropriate when aligning two segments of genomic DNA that may have local regions of similarity embedded in a background of a non-homologous sequence.

A Global Alignment. A global alignment is a sequence alignment over the entire length of two or more nucleic acid or protein sequences. In a global alignment the sequences are assumed to be homologous along their entire length [15].

3.5.1 Scoring systems in pairwise alignments

In order to align a pair of sequences, a scoring system is required to score matches and mismatches. The scoring system can be as simple as "+1" for a match and "-1" for a mismatch between the pair of sequences at any given site of comparison. However substitutions, insertions and deletions occur at different rates over evolutionary time. This variation in rates is the result of a large number of factors, including the mutation process, genetic drift and natural selection. For protein sequences, the relative rates of different substitutions can be empirically determined by comparing a large number of related sequences. These empirical measurements can then form the basis of a scoring system for aligning subsequent sequences. Many scoring systems have been developed in this way. These matrices incorporate the evolutionary preferences for certain substitutions over other kinds of substitutions in the form of log-odd scores. Popular matrices used for protein alignments are BLOSUM [9] and PAM [2] matrices.

Note: The BLOSUM matrix is a substitution matrix. The number of a BLOSUM matrix indicates the threshold (%) similarity between the sequences originally used to create the matrix. BLOSUM matrices with higher numbers are more suitable for aligning closely related sequences.

When aligning protein sequences in Geneious, a number of BLOSUM and PAM matrices are available.

3.5.2 Algorithms for pairwise alignments

Once a scoring system has been chosen, we need an algorithm to find the optimal alignment of two sequences. This is done by inserting gaps in order to maximize the alignment score. If the sequences are related along their entire sequence, a global alignment is appropriate. However, if the relatedness of the sequences is unknown or they are expected to share only small regions of similarity, (such as a common domain) then a local alignment is more appropriate.

An efficient algorithm for global alignment was described by Needleman and Wunsch [15], and their algorithms was later extended by Gotoh to model gaps more accurately [6]. For local alignments, the Smith-Waterman algorithm [19] is the most commonly used. See the references provided for further information on these algorithms.

3.5.3 Pairwise alignment in Geneious

A dotplot is a comparison of two sequences. A pairwise alignment is another such comparison with the aim of identifying which regions of two sequences are related by common ancestry and which regions of the sequences have been subjected to insertions, deletions, and substitutions.

Alignment options

The options available for the alignment cost matrix will depend on the kind of sequence.

- Protein sequences have a choice of PAM [2] and BLOSUM [9] matrices.
- Nucleotide sequences have choices for a pair of match/mismatch costs. Some scores distinguish between two types of mismatches: transition and transversion. Transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) generally occur more frequently than transversions. Differences in the ratio of transversions and transversions result in various models of substitution. When applicable, Geneious indicates the target sequence similarity for the alignment scores, i.e. the amount of similarity between the sequences for which those scores are optimal.
- Both protein and nucleotide pairwise alignments have choices for gap open / gap extension penalties/costs. Unlike many alignment programs these values are not restricted to integers in Geneious.

There is also a check box for "free gaps at ends". If this is selected, gaps at either end of the alignment are not penalized when determining the optimal alignment. This is especially useful if you are aligning sequence fragments that overlap slightly in their starting and ending positions (e.g. when using two slightly different primer pairs to extract related sequence fragments from different samples).

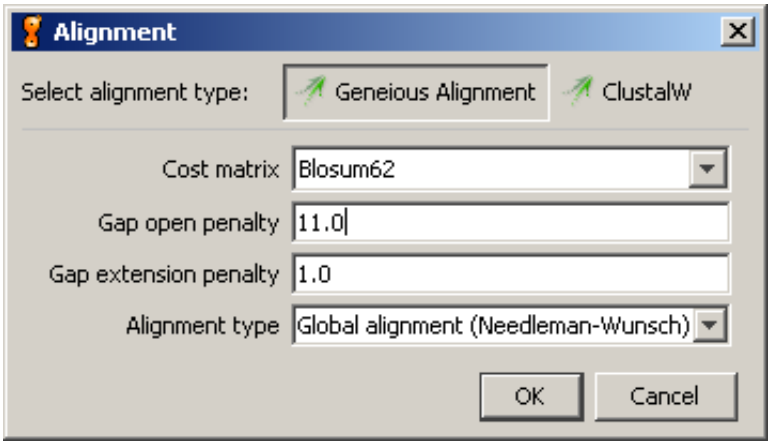


Figure 3.6: Options for protein pairwise alignment

3.6 Multiple sequence alignments

A multiple sequence alignment is a comparison of multiple related DNA or amino acid sequences. A multiple sequence alignment can be used for many purposes including inferring

the presence of ancestral relationships between the sequences. It should be noted that protein sequences that are structurally very similar can be evolutionarily distant. This is referred to as distant homology. While handling protein sequences, it is important to be able to tell what a multiple sequence alignment means – both structurally and evolutionarily. It is not always possible to clearly identify structurally or evolutionarily homologous positions and create a single "correct" multiple sequence alignment [3].

Multiple sequence alignments can be done by hand but this requires expert knowledge of molecular sequence evolution and experience in the field. Hence the need for automatic multiple sequence alignments based on objective criteria. One way to score such an alignment would be to use a probabilistic model of sequence evolution and select the alignment that is most probable given the model of evolution. While this is an attractive option there are no efficient algorithms for doing this currently available. However a number of useful heuristic algorithms for multiple sequence alignment do exist.

3.6.1 Progressive pairwise alignment methods

The most popular and time-efficient method of multiple sequence alignment is progressive pairwise alignment. The idea is very simple. At each step, a pairwise alignment is performed. In the first step, two sequences are selected and aligned. The pairwise alignment is added to the mix and the two sequences are removed. In subsequent steps, one of three things can happen:

- Another pair of sequences is aligned
- A sequence is aligned with one of the intermediate alignments
- A pair of intermediate alignments is aligned

This process is repeated until a single alignment containing all of the sequences remains. Feng & Doolittle were the first to describe progressive pairwise alignment [5]. Their algorithm used a guide tree to choose which pair of sequences/alignments to align at each step. Many variations of the progressive pairwise alignment algorithm exist, including the one used in the popular alignment software ClustalX [22].

3.6.2 Multiple sequence alignment in Geneious

Multiple sequence alignment in Geneious is done using progressive pairwise alignment. The neighbor-joining method of tree building is used to create the guide tree.

As progressive pairwise alignment proceeds via a series of pairwise alignments this function in Geneious has all the standard pairwise alignment options. In addition, Geneious also has the option of refining the multiple sequence alignment once it is done. "Refining" an alignment involves removing sequences from the alignment one at a time, and then realigning the removed

sequence to a profile of the remaining sequences. The number of times each sequence is re-aligned is determined by the "refinement iterations" option in the multiple alignment window. The resulting alignment is placed in the folder containing the sequences aligned.

In some cases building a guide tree can take a long time since it requires making a pairwise alignment between each pair of sequences. The "build guide tree via alignment" option may speed this part by taking a different route. First make a progressive multiple alignment using a random ordering, and use that alignment to build the guide tree. Notice that while this typically speeds up the process that may not be the case when the sequences are very distant genetically.

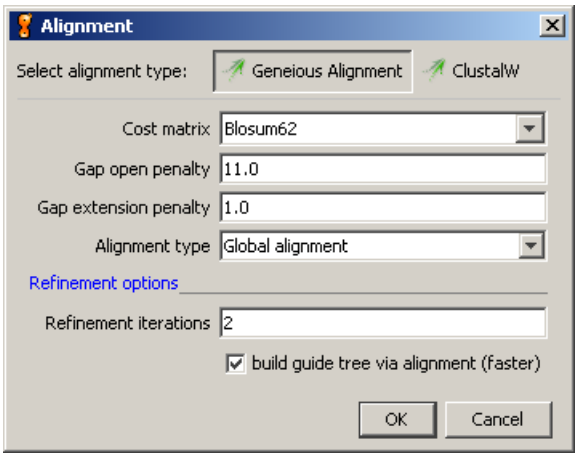


Figure 3.7: The multiple alignment window

3.7 Sequence alignment using ClustalW (*pro* only)

ClustalW is a widely used program for performing sequence alignment [23, 22]. If you have ClustalW installed your computer, *Geneious pro* allows you to run ClustalW directly from inside the program without having to export or import your sequences.

If you do not have ClustalW or are unsure if you do, you should attempt to perform a ClustalW alignment without specifying a location and Geneious will present you with options including details on how to download ClustalW and an automatic search for finding its location on your hard drive.

To perform an alignment using ClustalW, select the sequences or alignment you wish to align and select the "Alignment" button from the Toolbar. At the top of the alignment options window there are buttons allowing you to select the type of alignment you wish to do. Choose "ClustalW" here and the options available for a ClustalW alignment will be displayed.

The options are:

- *ClustalW Location:* This should be set to the location of the ClustalW program on your computer. Enter the path to it in the text field or click the "Browse" button to browse for the location. If the location is invalid and you attempt to perform an alignment Geneious will tell you and offer the options detailed above for getting or finding ClustalW.
- *Cost Matrix:* Use this to select the desired cost matrix for the alignment. The available options here will change according to the type of the sequences you wish to align. You can also click the "Custom File" button to use a cost matrix that you have on your computer (the format of these is the same as for the program BLAST).
- *Gap open cost* and *Gap extend cost:* Enter the desired gap costs for the alignment.
- *Free end gaps:* Select this option to avoid penalizing gaps at either end of the alignment. See details in the Pairwise Alignment section above.
- *Preserve original sequence order:* Select this option to have the order of the sequences in the table preserved so that the alignment contains the sequences in the same order.
- *Additional options:* Any additional parameters accepted by the ClustalW command line program can be entered here. Refer to the ClustalW manual for a description of the available parameters.

After entering the desired options click "OK" and ClustalW will be called to align the selected sequences or alignment. Once complete, a new alignment document will be generated with the result as detailed previously.

3.8 Building Phylogenetic trees

Geneious provides some basic phylogenetic tree reconstruction algorithms for a preliminary investigation of relationships between newly acquired sequences. For more sophisticated methods of phylogenetic reconstruction such as Maximum Likelihood and Bayesian MCMC we recommend specialist software such as PAUP* [20] and MRBAYES [17].

Geneious implements the Neighbor-joining [18] and UPGMA [14] methods of tree reconstruction.

3.8.1 Phylogenetic tree representation

A phylogenetic tree describes the evolutionary relationships amongst a set of sequences. They have a few commonly associated terms that are depicted in Figure 2.16 and are described below.

Branch length. A measure of the amount of divergence between two nodes in the tree. Branch lengths are usually expressed in units of substitutions per site of the sequence alignment.

Nodes or internal nodes of a tree represent the inferred common ancestors of the sequences that are grouped under them.

Tips or leaves of a tree represent the sequences used to construct the tree.

Taxonomic units. These can be species, genes or individuals associated with the tips of the tree.

A phylogenetic tree can be rooted or unrooted. A rooted tree consists of a root, or the common ancestor for all the taxonomic units of the tree. An unrooted tree is one that does not show the position of the root. An unrooted tree can be rooted by adding an outgroup (a species that is distantly related to all the taxonomic units in the tree). A common format for representing phylogenetic trees is the Newick format [12].

3.8.2 Neighbor-joining

In this method, neighbors are defined as a pair of leaves with one node connecting them. The principle of this method is to find pairs of leaves that minimize the total branch length at each stage of clustering, starting with a star-like tree. The branch lengths and an unrooted tree topology can quickly be obtained by using this method without assuming a molecular clock [18].

3.8.3 UPGMA

This clustering method is based on the assumption of a molecular clock [14]. It is appropriate only for a quick and dirty analysis when a rooted tree is needed and the rate of evolution is does not vary much across the branches of the tree.

3.8.4 Distance models or molecular evolution models for DNA sequences

The evolutionary distance between two DNA sequences can be determined under the assumption of a particular model of nucleotide substitution. The parameters of the substitution model define a rate matrix that can be used to calculate the probability of evolving from one base to another in a given period of time. This section briefly discusses some of the substitution models available in Geneious. Most models are variations of two sets of parameters – the *equilibrium frequencies* and *relative substitution rates*.

Equilibrium frequencies refer to the background probability of each of the four bases A, C, G, T in the DNA sequences. This is represented as a vector of four probabilities $\pi_A, \pi_C, \pi_G, \pi_T$ that sum to 1.

Relative substitution rates define the rate at which each of the transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) and transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, $G \leftrightarrow T$) occur in an evolving sequence. It is represented as a 4x4 matrix with rates for substitutions from every base to every other base.

Jukes Cantor

This is the simplest substitution model [10]. It assumes that all bases have the same equilibrium base frequency, i.e. each nucleotide base occurs with a frequency of 25% in DNA sequences and each amino acid occurs with a frequency of 5% in protein sequences. This model also assumes that all nucleotide substitutions occur at equal rates and all amino acid replacements occur at equal rates.

HKY

The HKY model [8] assumes every base has a different equilibrium base frequency, and also assumes that transitions evolve at a different rate to the transversions.

Tamura-Nei

This model also assumes different equilibrium base frequencies. In addition to distinguishing between transitions and transversions, it also allows the two types of transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) to have different rates [21].

3.8.5 Resampling – Bootstrapping and jackknifing

Resampling is a statistical technique where a procedure (such as phylogenetic tree building) is repeated on a series of data sets generated by sampling from an original data set. The results of analyzing the sampled data sets are then combined to generate summary information about the original data set.

In the context of tree building, resampling involves generating a series of sequence alignments by sampling columns from the original sequence alignment. Each of these alignments (known as pseudoreplicates) is then used to build an individual phylogenetic tree. A consensus tree is constructed from by combining information from the set of generated trees, providing an estimate for the level of support for each clade in the final consensus tree [4].

Bootstrapping is the statistical method of resampling with replacement. To apply bootstrapping in the context of tree building, each pseudo-replicate is constructed by randomly sampling columns of the original alignment with replacement until an alignment of the same size is obtained [4].

Jackknifing is a statistical method of numerical resampling based on deleting a portion of the original observations for each pseudo-replicate. A 50% jackknife randomly deletes half of the columns from the alignment to create each pseudo-replicate.

3.8.6 Tree building in Geneious

Geneious can build a phylogenetic tree for a set of sequences using pairwise genetic distances. To build a tree, select an alignment or a set of related sequences (all DNA or all protein) in the Document table and click the "Build Tree" icon or choose this option from the Tools menu.

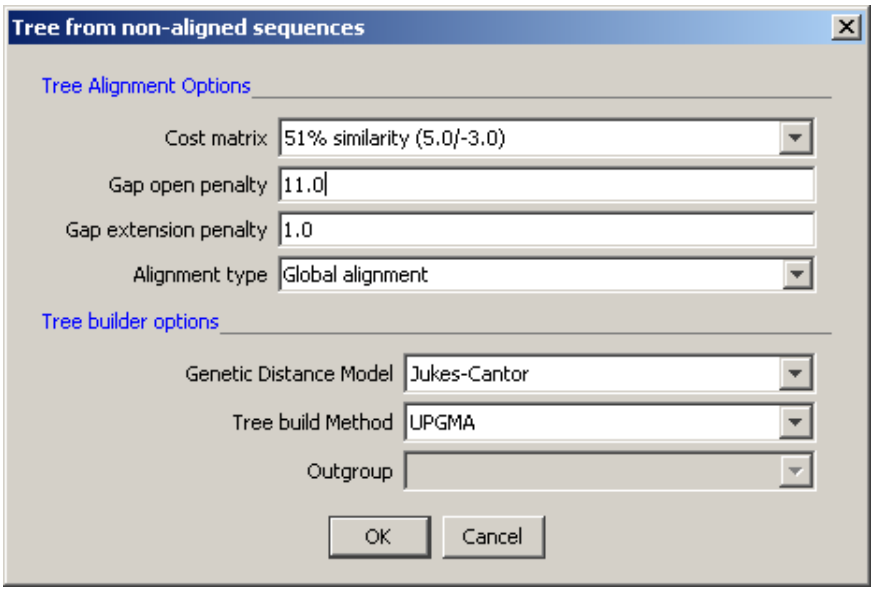


Figure 3.8: Tree building options in Geneious

If you are building a tree from an alignment, the following options are seen in the tree window.

Genetic distance model. This lets the user choose the kind of substitution model used to estimate branch lengths. If you are building a tree from DNA sequences you have the choices "Jukes Cantor", "HKY" and "Tamura Nei". If you are building a tree from amino acid sequences you only have the option of "Jukes Cantor" distance correction.

Tree building method. There are two methods under this option – Neighbor joining [18] and UPGMA [14].

Consensus method via resampling. Check this box to build a consensus tree using resampling of sequence alignment data.

Resampling method. Either bootstrapping or jackknifing can be performed when resampling columns of the sequence alignment.

Number of samples. The number of alignments and trees to generate while resampling and building a consensus tree. A value of at least 100 is recommended.

Support threshold. This is used to decide which monophyletic clades to include in the consensus tree, after comparing all the trees in the original set as defined below.

Resampling results in a set of trees, which we will refer to as the "original set of trees" for the definitions that follow.

A 100% support threshold results in a *"Strict consensus tree"* which is a tree where the included clades are those that are present in all the trees of the original set. A 50% threshold results in a *"Majority rule consensus tree"* that includes only those clades that are present in the majority of the trees in the original set. A threshold less that 50% gives rise to a *"Greedy consensus tree"*. In constructing a *"Greedy consensus tree"* clades are first ordered according to the number of times they appear (i.e. the amount of support they have), then the consensus tree is constructed progressively to include all those clades whose support is above the threshold **and** that are compatible with the tree constructed so far.

Note: The above definitions apply to rooted trees. The same principles can be applied to unrooted trees by replacing "clades" with "splits". Each branch (edge) in an unrooted tree corresponds to a different split of the taxa that label the leaves of this tree.

3.9 BLAST

BLAST (Basic Local Alignment Search Tool) is a fast method to search large databases of sequences for a sequences similar to a query sequence [1]. Geneious provides access to the NCBI sequence databases via BLAST. For more information on BLAST check out `http://www.ncbi.nlm.nih.gov/BLAST/`.

Data from BLAST are summarized in a way similar to protein and nucleotide sequences, and can be viewed in the same way too. However, data from BLAST have an extra e-value table and is a result of searching for a GenBank ID or a string of residues. The results of the BLAST are a list of the closest matches to the original query. Once you have a set of matches, you can use them to perform sequence alignments and build trees the way you do with nucleotide and protein sequences. For more detailed information on BLAST, refer to section 2.2.2 of this manual.

3.10 Results of analysis

All analysis results are deposited in the currently selected folder. If no local folder is selected then you will be prompted for a local folder. This applies to sequence alignments, phylogenetic trees, sequence translations, reverse complements and extraction of sequences. Once

generated, analysis results can be dragged to another location if desired.

Chapter 4

Collaboration (*pro* only)

Collaboration allows Geneious users to share the products of their research and work with each other. It uses an open internet protocol called 'Jabber' to move documents across the internet. It can work with any existing Jabber service, such as Google Talk, but we recommend using the Geneious default, talk.geneious.com.

If you have the knowledge to run and manage your own Jabber server, we recommend using Wildfire from Jive Software [http://www.jivesoftware.org/wildfire/] which is available for free under the GNU General Public License. [http://www.gnu.org/copyleft/gpl.html]

Collaboration is only available in *Geneious pro*.

This chapter shows you how to:

- Create a new collaboration account
- Search for, and add contacts to your account
- Share local folders with your contacts
- Search your contacts as you would an online database

4.1 Managing Your Accounts

When you start Geneious you will see the empty Collaboration service in the Services Panel and the Collaboration menu at the top. You can open the Add New Account dialog by either right clicking on Collaboration in the Services Panel and clicking, 'Add New Account' in the popup menu, or by selecting the same option from menu at the top.

4.1.1 Add New Account

In this dialog you are given the options of creating a new account on the server or entering the details for an existing account. In most cases you will be creating a new account, but there may be occasions when you need to re-enter the details of an existing account. If you choose to create a new account Geneious will attempt to automatically register your account on the server at the end of this process.

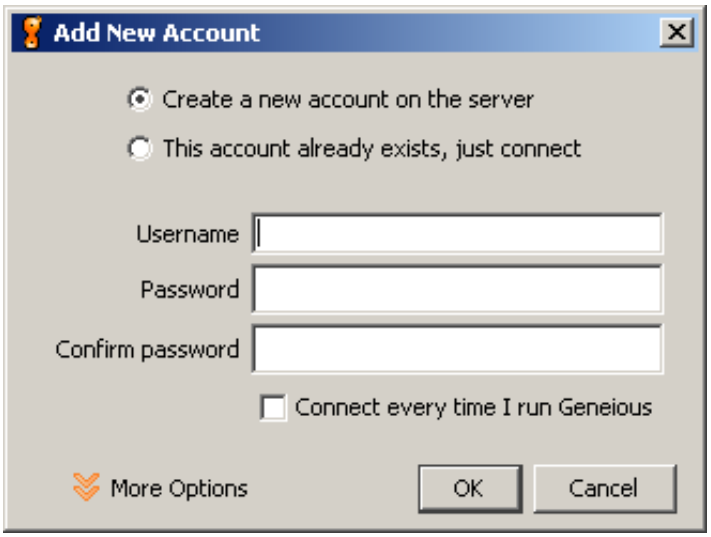


Figure 4.1: Add New Account dialog box

Choose a username and password now. Enter your password twice for a new account.

If you want this account to connect to the server every time you start Geneious, check the checkbox labelled, 'Connect every time I run Geneious'.

More Options

You can change some of the defaults for new and exiting accounts:

- *Account Name* is the name displayed in the Services Panel for this account. It defaults to your username if nothing is entered
- *Server* is the server your account connects to
- *Jabber Service Name* is required by some other Jabber service providers, such as Google Talk
- *Port Number* for Jabber servers running on a non-standard port.

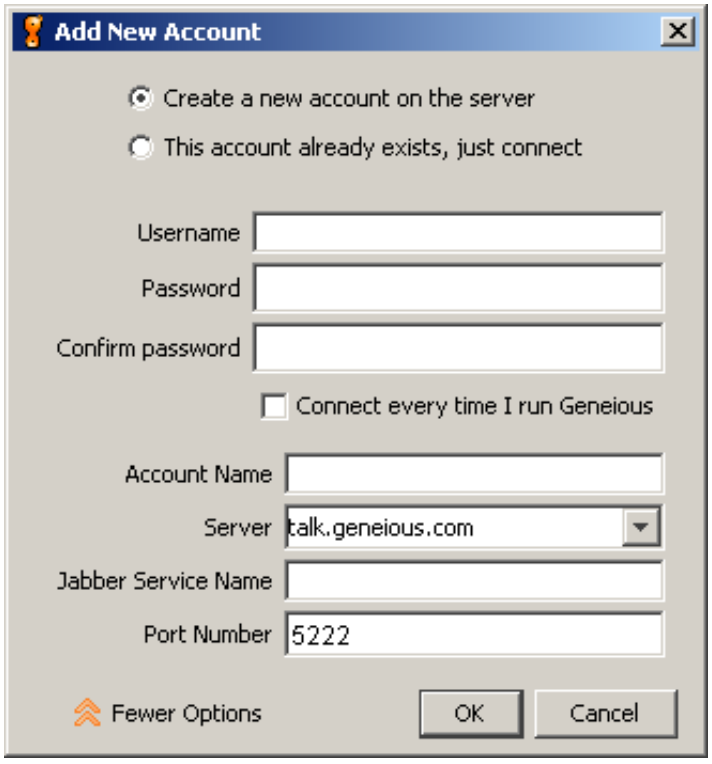


Figure 4.2: Add New Account dialog box with More Options

4.1.2 Edit Account Details

Select your account in the Services Panel and this option from the menu. Or right-click on your account and select it from the popup menu. This option is only available when your account is not connected.

This dialog has nearly all the same fields as the Add New Account dialog. Remember that if you change details such as your username or password you may not be able to connect to your account.

4.1.3 Connect/Disconnect

Connect to, or disconnect from, the server by right-clicking on your account or selecting your account in the Services Panel and choosing that option from the menu at the top.

4.1.4 Delete Account

You can delete an account from Geneious by right-clicking on your account or selecting your account in the Services Panel an choosing that option from the menu at the top.

This does not delete the account on the server.

4.2 Managing Your Contacts

Once you have an account and are connected you can start adding contacts. You will not be able to add contacts while an account is disconnected.

4.2.1 Add Contact

Select your account in the Services Panel and choose Add Contact from the menu at the top or right-click on your account in the Services Panel and choose the same option.

You will see a simple dialog with one field, Jabber ID. A Jabber ID looks like an email address and has a similar function: It uniquely identifies some other Geneious users account. You can enter a contact's Jabber ID directly into this field if you know it. To see your own Jabber ID hover your mouse over your account in the Services Panel and it will appear in a tool-tip.

If the server supports it, you should also see a 'Search For Contact' link. Click this to go to the next dialog.

Figure 4.3: Add Contact dialog box

Here you will see a box for a search string, and some checkboxes indicating what you are searching on. Enter all or part of the name or email of the contact you want and click the Search button. If any rows are returned in the results table you will be able to select one or entries and add them as contacts.

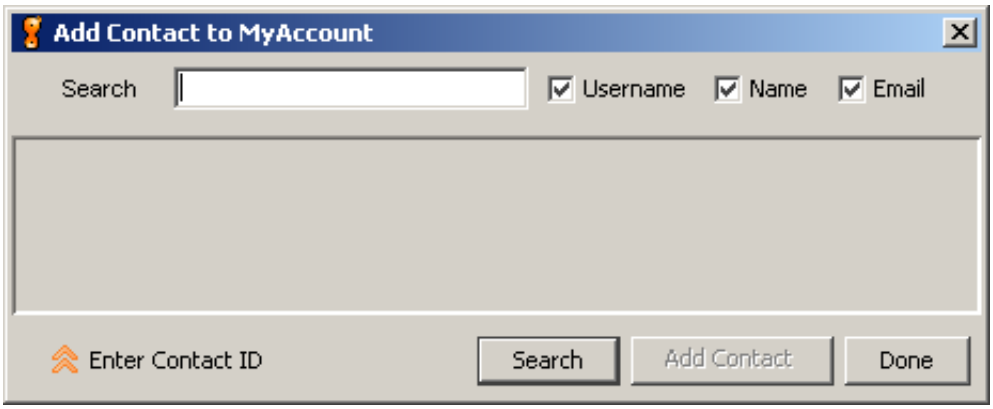


Figure 4.4: Add New Contact dialog box in searching mode

Your new contact will appear immediately in your contact list, however you will not be able to tell whether your new contact is online until they accept you as a contact. Similarly you will occasionally see a dialog box pop up asking you, 'Allow user.name@talk.geneious.com as contact?'

This is another Geneious user attempting to add you as a contact in this manner.

4.2.2 Rename Contact

You can rename a contact in your contact list by right-clicking the contact in the Services panel and selecting Rename Contact, then entering a new name for the contact.

This only changes the name displayed for the contact in the Services Panel.

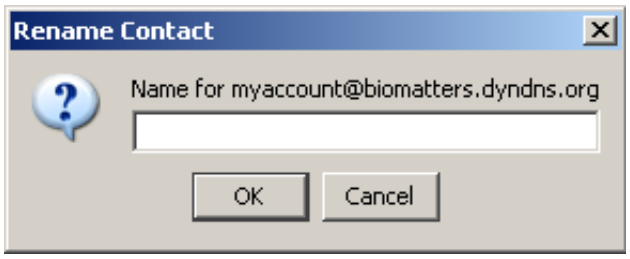


Figure 4.5: Rename Contact dialog box

4.2.3 Remove Contact

If you no longer wish to share documents with a contact, you can remove that contact by right-clicking the contact in the Services panel and selecting "Remove Contact". This deletes you from their contact list as well. If you find that a contact has disappeared from your list, this may be the reason.

4.3 Sharing Documents

Select one of your local folders. Select Share Folder from the File menu. Alternatively right-click on a local folder and select the same option.

- If you share a folder all documents in that folder are shared.
- If you share a folder all sub-folders of that folder are shared.
- If you share a folder it is available to all your contacts.

4.4 Searching Contacts

You can search each of your contacts individually, or you can search all the contacts in an account. The process is the same in each case.

Select an online account or an online contact. You will see a search panel just as you would for an online database. If you search an account, you will retrieve documents from all contacts who are online. If you search just a contact you will retrieve documents from that contact alone. You can also set up agents just as you would for any database. The agents will fail if they try to search when an account of a contact is offline though.

NOTE: The search panel is still enabled and you can still click the Search button if a contact is offline. This is so that documents you have downloaded are still available if a contact suddenly goes offline. Any attempt to search that contact will fail of course.

4.4.1 Concurrency

The current iteration of Collaboration is somewhat restricted in that it can only send one message to the server at a time. This means that if you search several contacts at once, you are really searching them one after the other.

Additionally when someone downloads a document from you, it is actually your copy of Geneious sending a message to theirs, and it must complete this before it attempts the next search.

If someone is downloading a large document from you, you will see the text '(busy)' next to your account name. Searches will not complete while your account is busy. If you really need to do a search, and you don't mind interrupting someone else's download, you can disconnect and reconnect your account. This will reset your connection to the server.

Future editions of Geneious will not have this restriction.

Bibliography

[1] SF. Altschul, W. Gish, W. Miller, EW. Myers, and DJ. Lipman, *Basic local alignment search tool.*, J Mol Biol **215** (1990), no. 3, 403–410. 15, 20, 32, 35, 67

[2] MO. Dayhoff (ed.), *Atlas of protein sequence and structure*, vol. 5, National biomedical research foundation Washington DC, 1978. 59, 60

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge University Press, 1998. 61

[4] J. Felsenstein, *Confidence limits on phylogenies: An approach using the bootstrap.*, Evolution **39** (1985), no. 4, 783–791. 65

[5] DF. Feng and RF. Doolittle, *Progressive sequence alignment as a prerequisite to correct phylogenetic trees.*, J Mol Evol **25** (1987), no. 4, 351–60. 61

[6] O. Gotoh, *An improved algorithm for matching biological sequences.*, J Mol Biol **162** (1982), 705–708. 59

[7] M. Vingron HA. Schmidt, K. Strimmer and A. von Haeseler, *Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing.*, Bioinformatics **18** (2002), no. 3, 502–504. 16

[8] M. Hasegawa, H. Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial dna.*, J Mol Evol **22** (1985), no. 2, 160–174. 65

[9] S. Henikoff and JG. Henikoff, *Amino acid substitution matrices from protein blocks.*, Proc Natl Acad Sci U S A **89** (1992), no. 22, 10915–10919. 59, 60

[10] T. Jukes and C. Cantor, *Evolution of protein molecules*, pp. 21–32, Academic Press, New York, 1969. 65

[11] S. Kumar, K. Tamura, and M. Nei, *Mega3: Integrated software for molecular evolutionary genetics analysis and sequence alignment.*, Brief Bioinform **5** (2004), no. 2, 150–163. 18

[12] DR. Maddison, DL. Swofford, and WP. Maddison, *Nexus: an extensible file format for systematic information.*, Syst Biol **46** (1997), no. 4, 590–621. 16, 18, 64

[13] JV. Maizel and RP. Lenk, *Enhanced graphic matrix analysis of nucleic acid and protein sequences.*, Proc Natl Acad Sci U S A **78** (1981), no. 12, 7665–9. 57, 58

[14] C. Michener and R. Sokal, *A quantitative approach to a problem in classification.*, Evolution **11** (1957), 130–162. 63, 64, 66

[15] SB. Needleman and CD. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.*, J Mol Biol **48** (1970), no. 3, 443–53. 58, 59

[16] C. Notredame, DG. Higgins, and J. Heringa, *T-coffee: A novel method for fast and accurate multiple sequence alignment.*, J Mol Biol **302** (2000), no. 1, 205–217. 15

[17] F. Ronquist and JP. Huelsenbeck, *Mrbayes 3: Bayesian phylogenetic inference under mixed models.*, Bioinformatics **19** (2003), no. 12, 1572–4. 63

[18] N. Saitou and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.*, Mol Biol Evol **4** (1987), no. 4, 406–25. 63, 64, 66

[19] TF. Smith and MS. Waterman, *Identification of common molecular subsequences*, Journal of Molecular Biology **147** (1981), 195–197. 58, 59

[20] DL. Swofford, *Paup*: Phylogenetic analysis using parsimony (*and other methods) version 4.0*, Sinauer Assoc., Sunderland, Mass., 1999. 63

[21] K. Tamura and M. Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees.*, Mol Biol Evol **10** (1993), no. 3, 512–526. 65

[22] JD. Thompson, TJ. Gibson, F. Plewniak, F. Jeanmougin, and DG. Higgins, *The clustal x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.*, Nucleic Acids Res **25** (1997), no. 24, 4876–4882. 14, 15, 16, 61, 62

[23] JD. Thompson, DG. Higgins, and TJ. Gibson, *Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.*, Nucleic Acids Res **22** (1994), no. 22, 4673–4680. 14, 16, 62